Partek® Express™

Copyright

Copyright 1993-2010 by Partek Incorporated.

The software described in this document is furnished under a license agreement. The software may be used or copied only under the terms of the license agreement. No part of this manual may be reproduced in any form without prior written consent from Partek Incorporated. Licensee shall prevent and not permit any third parties, persons, or entities from copying, reproducing, duplicating, examining, inspecting, studying, and/or reviewing the Software, Documentation, and/or Information.

Partek, Partek Pro, Partek Express, Partek Analytical Spreadsheet, and Pattern Visualization System are registered trademarks of Partek Incorporated. All other brand and product names mentioned are trademarks owned by their respective companies or organizations.

Copyright © 2010 Partek Incorporated. St. Louis, MO.

Partek® Express™: Get Started with Partek® Express™

Partek Express

Partek® Express™ is a stand-alone software package that is produced by Partek Incorporated and is distributed by Affymetrix Incorporated. This chapter will briefly introduce each chapter in the Partek Express documentation.

Chapter 1: Get Started with Partek® Express™

Chapter 1 explains the chapters in Partek® Express™ Users Manual.

Chapter 2: Installation

Chapter 2 explains system recommendations, installation instructions, and library file directory set-up for Partek® Express™.

Chapter 3: The Guided Workflow

Chapter 3 explains the intuitive guided workflow in Partek® Express™.

Chapter 4: Create a New Study

Chapter 4 describes how to create a new study in Partek® Express™.

Chapter 5: Edit Sample Information

Chapter 5 describes how to edit sample information with Partek® Express™.

Chapter 6: Data Import

Chapter 6 explains the algorithm, library files, and annotation files used when importing .CEL or .CHP files.

Chapter 7: Quality Assessment

Chapter 7 describes quality assessment procedures in Partek® Express™, such as generating quality control metrics, invoking quality control plots, and applying quality control checks. Detailed information about the quality control metrics is also explained in this chapter.

Chapter 8: Principle Components Analysis

Chapter 8 explains the principle components analysis plot, as well as its function, in Partek® Express™.

Chapter 9: Effect Sizes

Chapter 9 explains the use and function of effect size in experimental design. Partek® Express™ contains plots to give a visual representation of effect size.

Chapter 10: Gene Significance

Chapter 10 explains how to view gene significance results in Partek® Express™.

Chapter 11: Power Analysis

Chapter 11 describes the benefits and how to use power analysis in Partek® Express™.

Chapter 12: Report

Chapter 12 breaks down the workflow report in Partek® Express™.

Chapter 13: Pathway Analysis

Chapter 13 describes how to invoke external Pathway Analysis software from Partek® Express™.

Chapter 14: Menu

Chapter 14 describes the menu shortcuts and functions in Partek® Express™.

Partek® Express™: Installation

Partek® Express™ is a stand-alone software package. It can be installed on a Windows® computer with or without other software packages like Affymetrix® GCOS™, Affymetrix® Expression Console™, and/or Partek® Genomic Suite™.

System Recommendations

- Operating System: Windows® XP SP3, Vista, or Windows 7; Linux; Mac
- Operating System Language: English*
- CPU: 2 GHz or higher
- Memory (RAM): 1 GB or higher
- Free disk space: 20 GB or higher
- Monitor: 800x600 resolution or higher
- Mouse: 2 button with scroll wheel
- Internet connection for the purpose of downloading library and annotation files
- Web browser with the Adobe® PDF plug-in for the purpose of viewing the User's Manual
- *Partek Express can run on non-English versions of Windows, but there are restrictions. For example, you should be able to finish the on-line tutorial (http://www.partek.com/Tutorials), but CEL file names and sample attributions need to be in English.

Installation Instructions

The Partek Express download page is located at: http://www.partek.com/html/updates.html.

- Download and run the installer. The web browser may pop up a **Security Warning** dialog like the following:

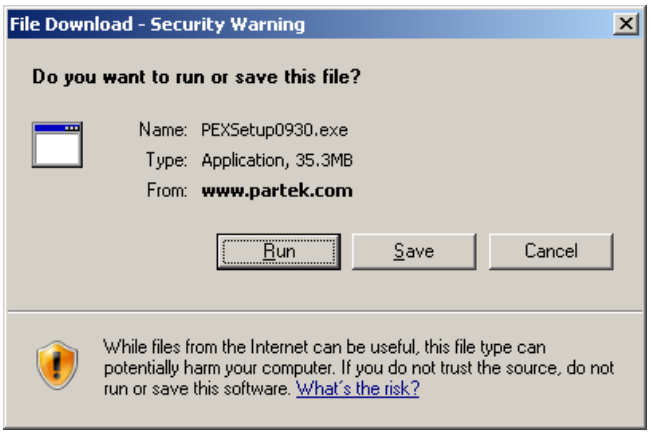

File Download - Security Warning
Do you want to run or save this file?
Name: PEXSetup0930.exe
Type: Application, 35.3MB
From: www.partek.com
Run
Save
Cancel
While files from the Internet can be useful, this file type can potentially harm your computer. If you do not trust the source, do not run or save this software. What's the risk?

Figure 2. 1: Download and Run the Installer from a Web Browser

- Select **Run** to download and run the installer
- After the downloading is finished, the **Partek Express Setup Wizard** will appear (Figure 2. 2). Select **Next >**

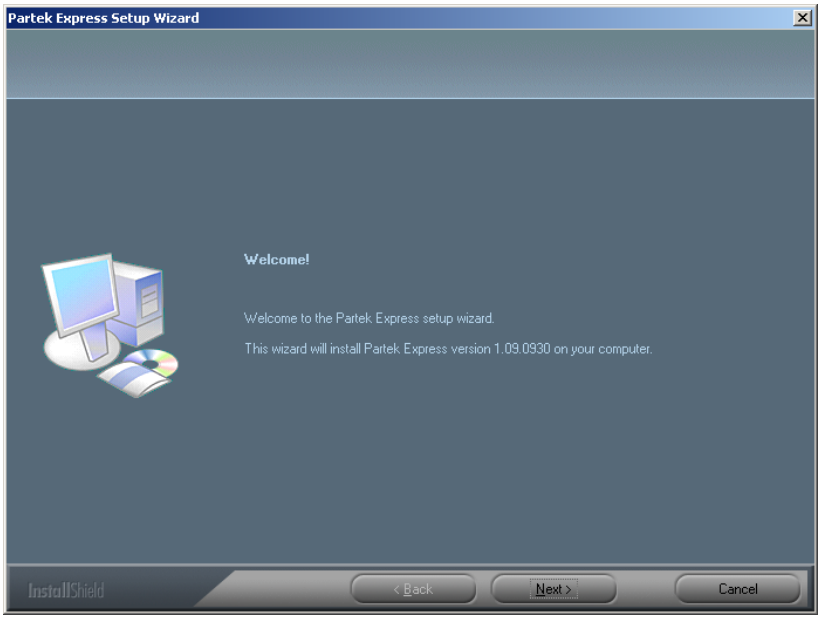

Partek Express Setup Wizard
Welcome!
Welcome to the Partek Express setup wizard.
This wizard will install Partek Express version 1.09.0930 on your computer.
InstallShield
< Back
Next >
Cancel

Figure 2. 2: Install Partek Express

- In the *Choose Installation Directory* step, you can select the install location (Figure 2. 3). By default, Partek Express will be installed in *C:\Program Files\Partek Express*. To change the default location, select **Browse…**. Select **Next >**

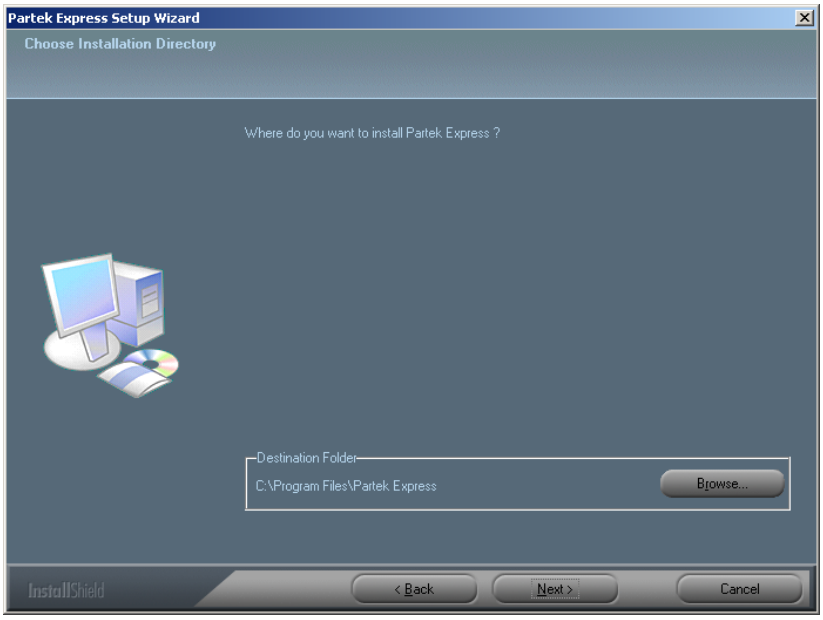


Figure 2. 3: Choose Installation Directory

- To import Affymetrix .CEL/.CHP files, Partek Express will need to download and store library/annotation files, so in the *Choose Microarray Library Directory* step (Figure 2. 4), you will need to select the location for downloading and storing those files. The default folder is *C:/Microarray Libraries*, to change the default folder, select **Browse…**. Make sure you have enough disk space and have read and write permission to the directory you select.
- Note: if the same library directory is also specified in Affymetrix® Expression Console™ or Partek® Genomic Suite™, Partek Express will share the library and annotation files with these software packages. Please refer to the corresponding user's guide on how to specify library directory in those software packages.
- Select **Next >**

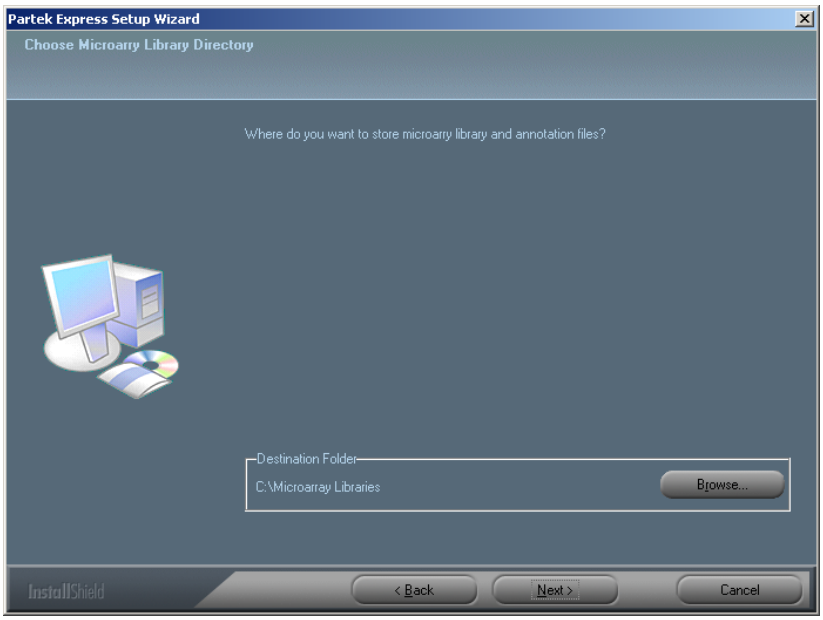


Figure 2. 4: Choose Microarray Library Directory

- Review the installation options (Figure 2. 5). Select **Next >**

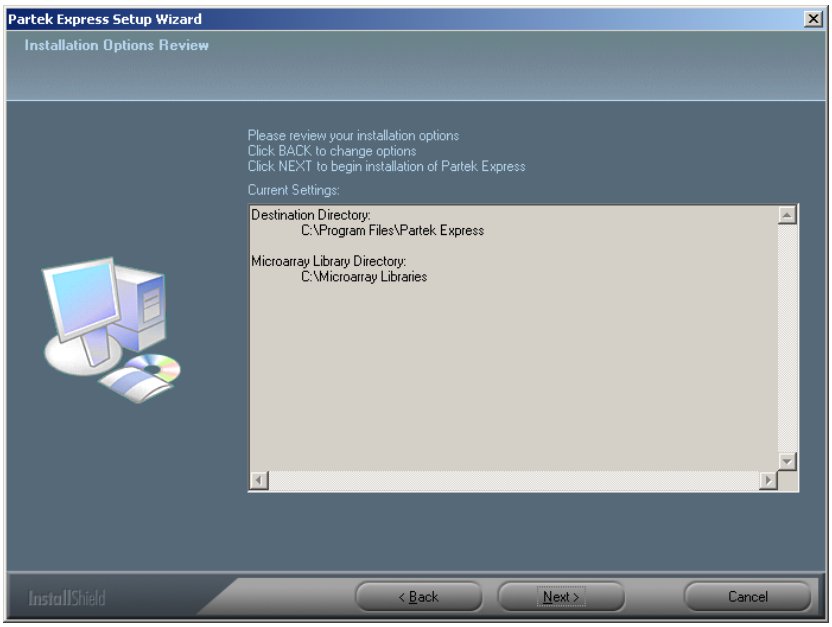


Figure 2. 5: Installation Options Review

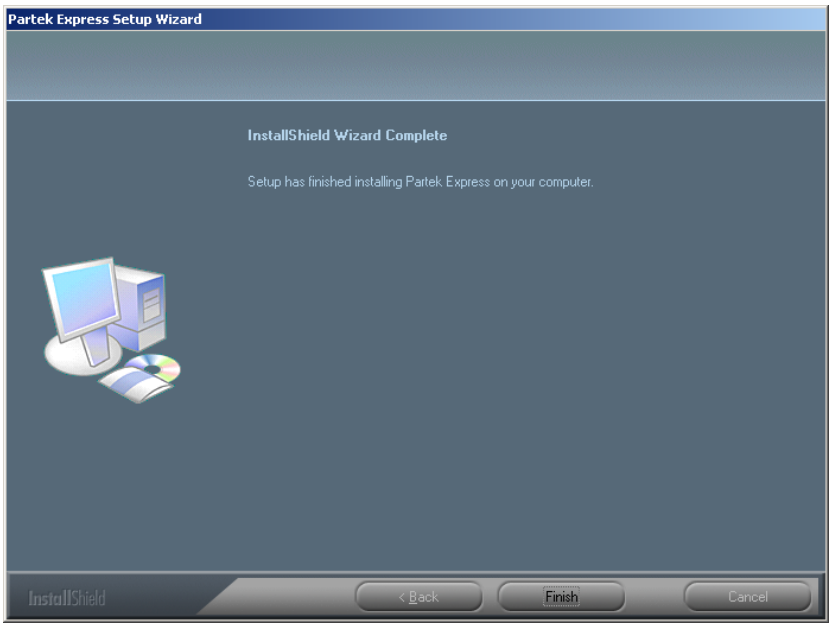


Figure 2. 6: Installation Complete

- Select the **Finish** button to complete the installation

Node-Locked License Setup

Upon successful installation, double-click the Partek Express icon on the desktop to launch the software. If a license could not be obtained, a *Partek Express Initialization Error* dialog will show up (Figure 2. 7).

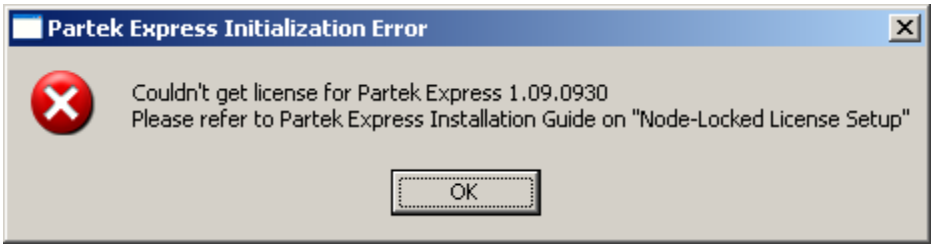

Partek Express Initialization Error
Couldn't get license for Partek Express 1.09.0930
Please refer to Partek Express Installation Guide on "Node-Locked License Setup"
OK

Figure 2. 7: Setup License Error

To request a node-locked license:

- Select the *Windows* **Start** menu **> All Programs > Partek > Partek Express > lmtools >** select the **System Settings** tab (Figure 2. 8)

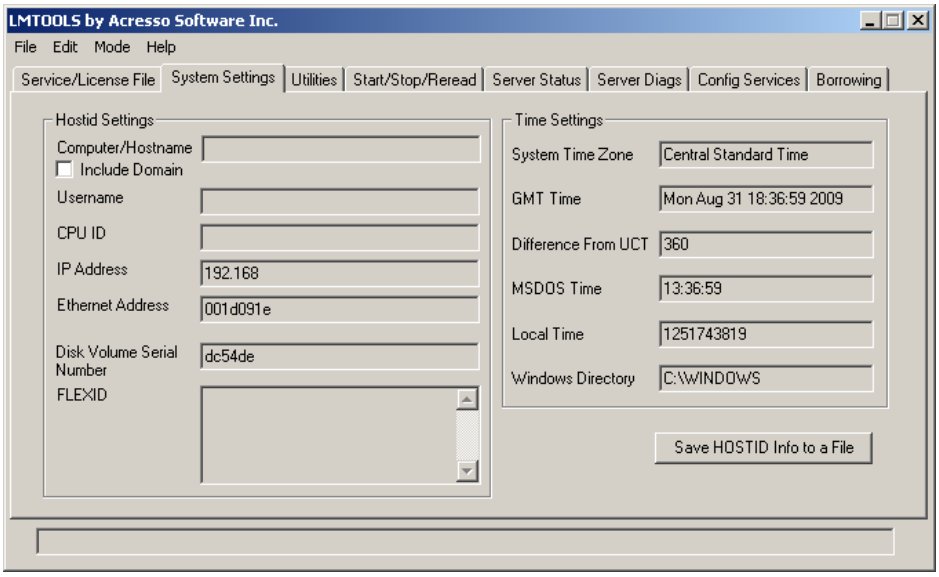


Figure 2. 8: LMTOOLS System Settings

- Select the **Save HOSTID Info to a File** button to save the system information to a file, such as *mySystemInfo.txt*
- Send an email to licensing@partek.com with the following information:
- Your name
- Your institution
- Institution address
- Phone number
- Partek product name with version number: Partek Express x.xx.xxxx
- Note: You can find this information from the Updates page:
  http://www.partek.com/html/updates.html
- Attach the *mySystemInfo.txt* saved by LMTOOLS
- The licensing request should be from an official email address (not gmail, hotmail, etc.)
- After getting a license file from Partek, save the file to **C:\Program Files\Partek Express\license\license.dat**

- If you specified a different installation directory (Figure 2. 3), right click the Partek Express icon on the desktop then choose *Properties* to find your installation directory
- Launch Partek Express again

Library Directory Setup

This section is only necessary if the installer has failed to create a *Microarray Library Directory* as shown in Figure 2. 4. Otherwise, upon the first launch of the software, Partek Express will prompt you to specify a library directory:

Figure 2. 9: Library File Directory

- Use *Windows Explorer* to create a directory with *read* and *write* permissions (e.g. C:\Microarray Libraries)

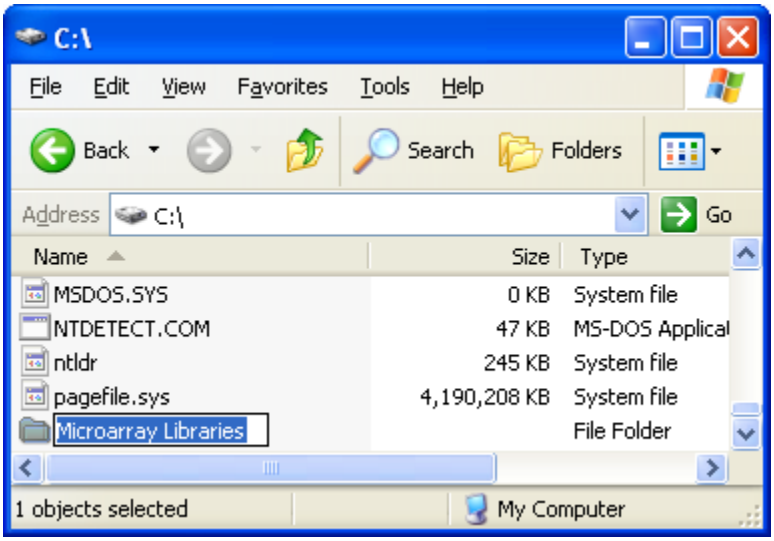


Figure 2. 10: Creating a directory with read and write permissions

- Return to Partek Express and select **OK** (Figure 2. 9)
- Specify the newly created directory (e.g., C:\Microarray Libraries) and select **OK**

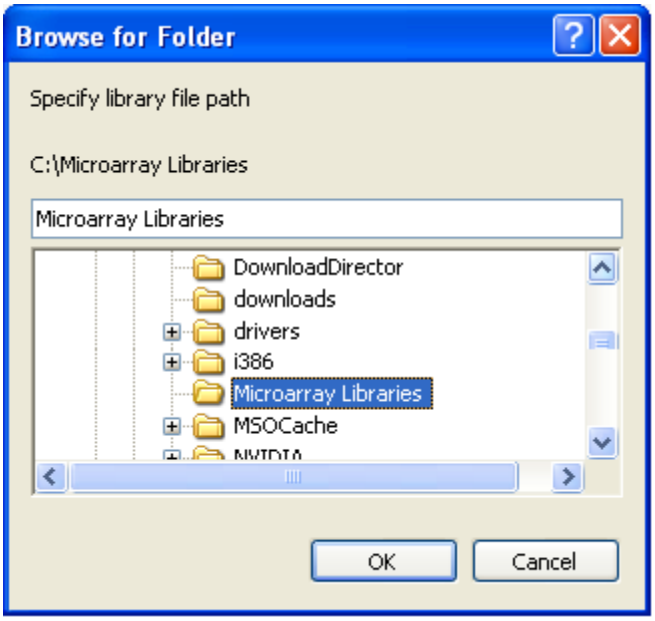


Figure 2. 11: Specify the library directory

Note: if the same library directory is also specified in Affymetrix® Expression Console™ or Partek® Genomic Suite™, Partek Express will share the library and annotation files with these software packages. Please refer to the corresponding user's guides on how to specify library directory in those software packages.

Partek® Express™: The Guided Workflow

Partek® Express™ Workflow

Partek® Express™ provides a guided workflow (Figure 3. 1), which is divided into individual steps for study definition, data import, quality control, principle data analysis, and effect sizes estimate. The workflow is designed to be intuitive through the use of dialogs. Simply follow the instructions in the current dialog to get to the next step. You can choose to stop at any step, save your study, and exit. When the saved study is loaded back into Partek Express, you can resume from your last step.

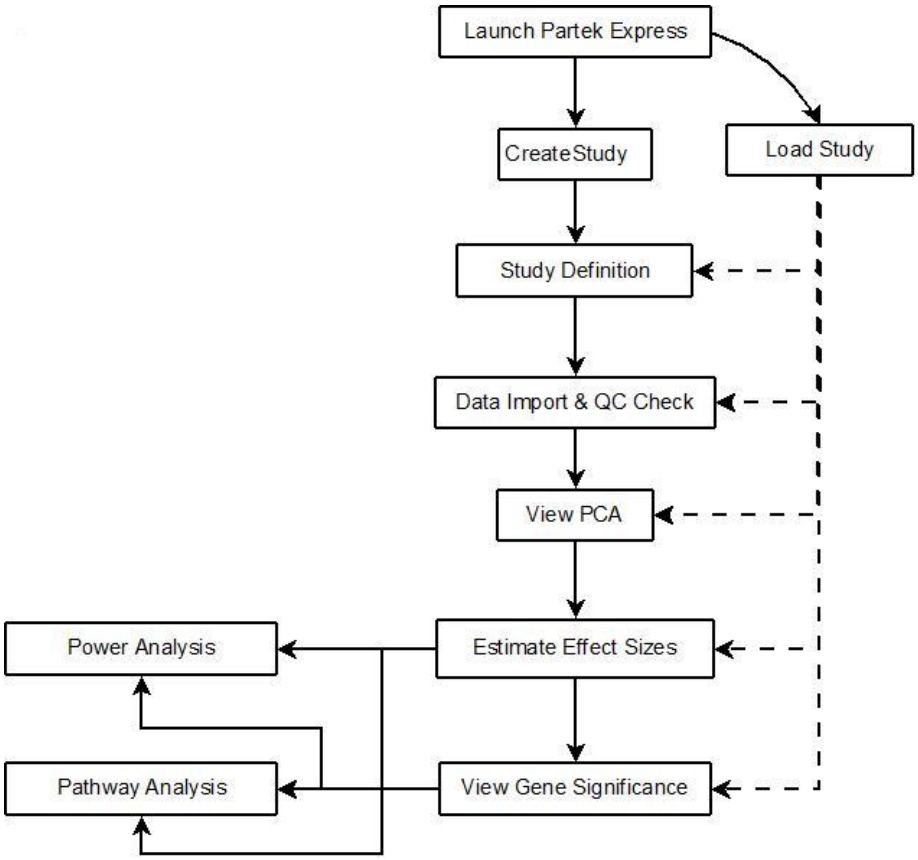


Figure 3. 1: Partek Express Work Flow Diagram

The example study shown in Figure 3. 2 has reached the *Effect Sizes* step. The main screen parts are:

1. The current study name
2. The *Menu* bar
3. The *Tools* bar
4. The *Result* and *Report* tabs. Each step produces its own result tab. You can select a tab to view the result from a previous step. The *Report* tab shows the information of the current study
5. The results of current step
6. The *Tell Me More…* button; selecting it will bring up the corresponding documentation chapter for the current step
7. Description of the next step
8. Optional workflow steps
9. The *Back* button; selecting it will delete the current step's result and go back one step. The difference between the *Back* button and the *Results* tab is that the Back button will destroy the current step's result and can only go back one step at a time; however, the *Results* tab can be used to view results (including the Report) from other steps.
10. The *Next* button; selecting it will proceed to the next step and eventually finish the workflow
11. The *Progress Bar*
12. Finished steps are colored in light red
13. The current step is highlighted in **Bold**
14. Future steps are colored in grey

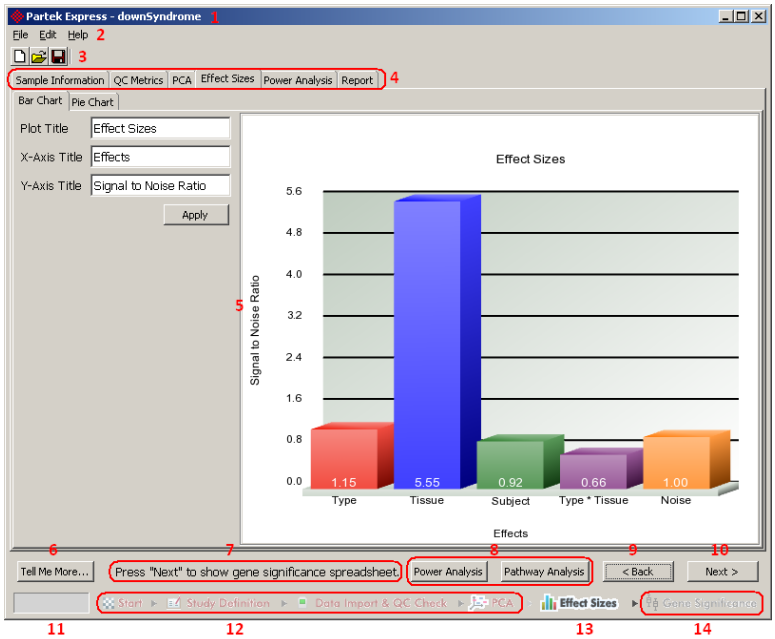


Figure 3. 2: An Example Study at the Effect Sizes Step

Partek® Express™: Creating a New Study

Introduction

This chapter describes how to create a new study in Partek® Express™.

Specify the Study File

A new study can be created by doing any of the following:

- Selecting **File > New Study**
- Selecting the *New Study* icon "🗋"
- Selecting the **Create Study** button at the bottom right of the main screen

A file browser will appear to let you specify a location and file name for the new study (Figure 4. 1). Give the *File* name, then select the **Save** button.

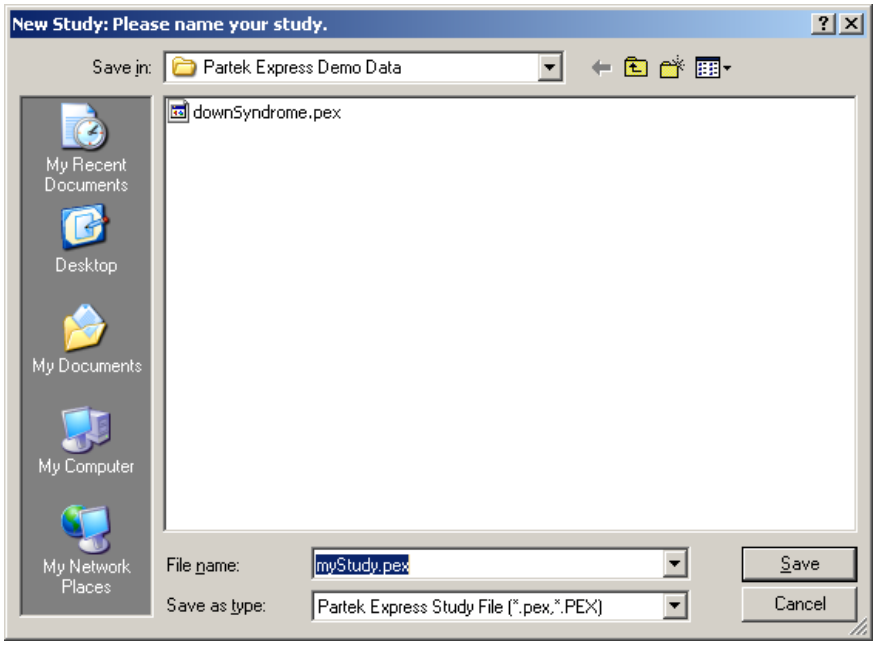


Figure 4. 1: Specify the Study File

Select Samples

After the study file is created, a sample selector window will appear so that you can navigate through folders to select samples for the study (Figure 4. 2). The main screen parts are:

1. The current directory; you can change the directory by typing in or pasting a new address
2. The *Go Back* icon; selecting this will take you to the previous directory
3. The *Go Up* icon; selecting this will take you up one level in the directory
4. The *Browse…* button; selecting this will bring up the system directory browser
5. The main directory browser; from here, select the directory where the sample .CEL/.CHP files are located
6. .CEL/.CHP files in the current directory will show in each row. If there are associated .ARR files, the sample attributes will also be shown here.
7. Selecting the column header's white box will Select/Deselect all samples
8. Selecting the white box will Select/Deselect individual or highlighted samples
9. <Click>, <Control+Click>, or <Shift+Click> to highlight samples
10. Selecting the corner square will de-highlight all samples
11. The *Add Samples* button will add selected samples and close the window

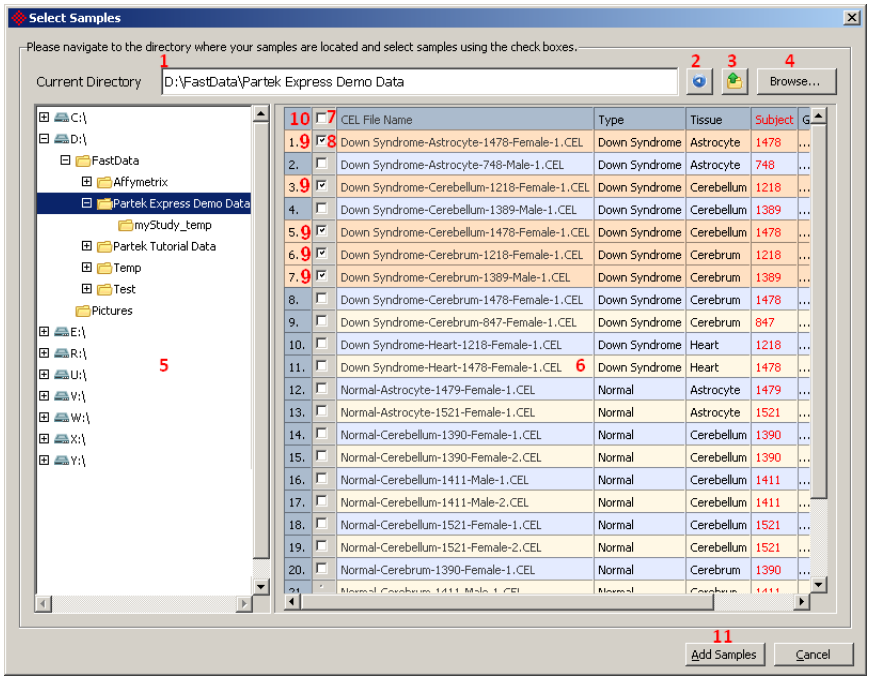


Figure 4. 2: Viewing the sample selector

Only files from the same folder can be added at the same time. After adding samples from a folder, select the **Add More Samples** icon " " on the *Sample Information Editor* tool bar to add samples from a different folder.

Partek® Express™: Edit Sample Information

Partek® Express™ Sample Editor

This chapter describes how to edit sample information using Partek® Express™. The sample editor shows sample attributes in a tabular format (Figure 5. 1). Each row corresponds to one sample. Each column corresponds to one sample attribute. If the ARR files are located for intensity or summarization files, the sample attribute information stored in them will be automatically extracted and filled into the sample editing table. You can then use the sample editor to add more attributes, delete unnecessary attributes, rearrange samples or sample attribute order. The rest of this chapter will provide a detailed guide on how this can be done.

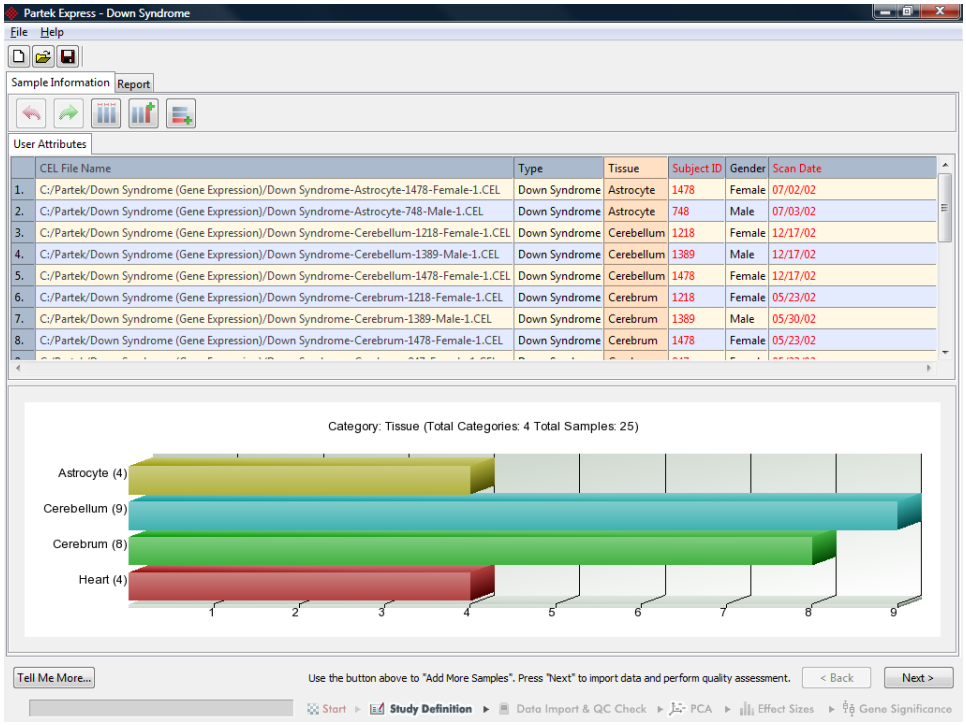

Partek Express - Down Syndrome
File Help
Sample Information Report
User Attributes
CEL File Name Type Tissue Subject ID Gender Scan Date
1. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Astrocyte-1478-Female-1.CEL Down Syndrome Astrocyte 1478 Female 07/02/02
2. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Astrocyte-748-Male-1.CEL Down Syndrome Astrocyte 748 Male 07/03/02
3. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Cerebellum-1218-Female-1.CEL Down Syndrome Cerebellum 1218 Female 12/17/02
4. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Cerebellum-1389-Male-1.CEL Down Syndrome Cerebellum 1389 Male 12/17/02
5. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Cerebellum-1478-Female-1.CEL Down Syndrome Cerebellum 1478 Female 12/17/02
6. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Cerebrum-1218-Female-1.CEL Down Syndrome Cerebrum 1218 Female 05/23/02
7. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Cerebrum-1389-Male-1.CEL Down Syndrome Cerebrum 1389 Male 05/30/02
8. C:/Partek/Down Syndrome (Gene Expression)/Down Syndrome-Cerebrum-1478-Female-1.CEL Down Syndrome Cerebrum 1478 Female 05/23/02
Category: Tissue (Total Categories: 4 Total Samples: 25)
Astrocyte (4)
Cerebellum (9)
Cerebrum (8)
Heart (4)
1 2 3 4 5 6 7 8 9
Tell Me More...
Use the button above to "Add More Samples". Press "Next" to import data and perform quality assessment.
< Back
Next >
Start Study Definition Data Import & QC Check PCA Effect Sizes Gene Significance

Figure 5. 1: Viewing the Sample Information editor

Getting Started

Column Types

The three types of columns in the Partek Express sample editor are listed below.

Type	Description
text	variable length string
numeric	double precision floating point (8 bytes) (-1.7E308 to 1.7E308)
categorical	Variable length nominal

Table 5. 1: Reviewing the column types in Partek Express

Different column types are displayed in different colors. Text columns are shown in gray; numeric columns are shown in blue; categorical columns with random effect are shown in red; and categorical columns with fixed effect are shown in black (see below for more details on random and fixed effects).

Random Effect vs. Fixed Effect

In a study, if an effect has all possible categories of interest, then it's usually a fixed effect. Effects like gender, disease state, tissue type, and treatment are usually fixed effects. Effects like subject, batch, and operator are usually random effects. In addition, if an effect has a subset of all possible categories of interest, then it's usually a random effect.

For example, if a study has 10 patients. They represent only a random sample of the global subjects of which an inference is being made. This patient effect is a random effect.

Here is another way to tell if a factor is random or fixed: imagine repeating the study. Would the same categories of each effect be used again?

Gender - Yes, the same genders would be used again - a fixed effect
Subject - No, the samples would be taken from other subjects - a random effect

You can specify an effect to be random or fixed by right-clicking on the column header then selecting **Properties**.

Note: Specifying an effect as a fixed or random effect will affect future statistical results. Please consult a statistician if you are not sure whether an effect is fixed or random.

View Sample Distribution for a Categorical Column

To view the sample attribute distribution for a column, single click on the header of a categorical column. The pane at the bottom of the sample editor will show a bar chart of sample attribute values (Figure 5. 2). Note: the sample distribution bar chart is only available for categorical columns, when clicking on the *text or a numeric column*, the bar chart will be blank. The main screen parts are:

1. Column header; selecting it will show the sample distribution

2. The category name (column header as in 1.)
3. The total number of categories in a column
4. The total number of samples
5. The individual category and the frequency in parentheses

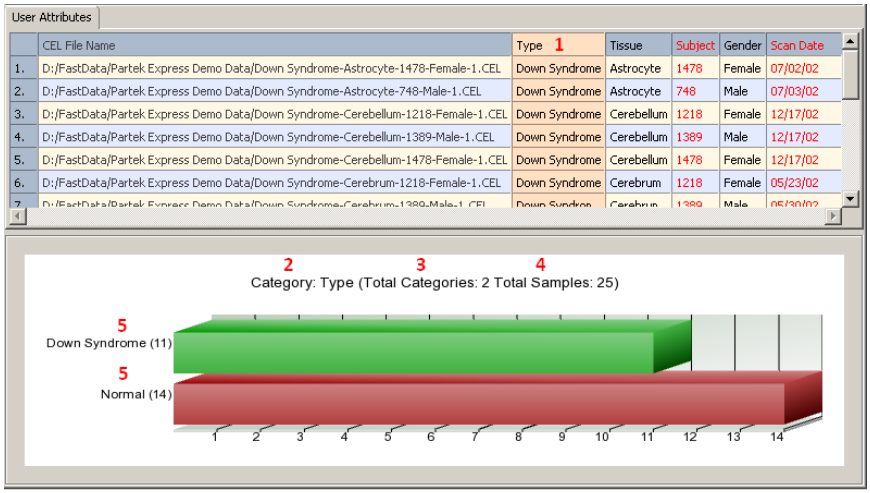

User Attributes
CEL File Name
Type 1
Tissue
Subject
Gender
Scan Date
1. D:/FastData/Partek Express Demo Data/Down Syndrome-Astrocyte-1478-Female-1.CEL Down Syndrome Astrocyte 1478 Female 07/02/02
2. D:/FastData/Partek Express Demo Data/Down Syndrome-Astrocyte-748-Male-1.CEL Down Syndrome Astrocyte 748 Male 07/03/02
3. D:/FastData/Partek Express Demo Data/Down Syndrome-Cerebellum-1218-Female-1.CEL Down Syndrome Cerebellum 1218 Female 12/17/02
4. D:/FastData/Partek Express Demo Data/Down Syndrome-Cerebellum-1389-Male-1.CEL Down Syndrome Cerebellum 1389 Male 12/17/02
5. D:/FastData/Partek Express Demo Data/Down Syndrome-Cerebellum-1478-Female-1.CEL Down Syndrome Cerebellum 1478 Female 12/17/02
6. D:/FastData/Partek Express Demo Data/Down Syndrome-Cerebrum-1218-Female-1.CEL Down Syndrome Cerebrum 1218 Female 05/23/02
2
3
4
Category: Type (Total Categories: 2 Total Samples: 25)
5
Down Syndrome (11)
5
Normal (14)
1 2 3 4 5 6 7 8 9 10 11 12 13 14

Figure 5. 2: View Sample Distribution for a Categorical Column

Read-only Columns

In order to maintain data integrity for your study, some columns are read only. For example, in Figure 5. 2, the column *CEL File Name* and *Scan Date* are read-only columns. If you attempt to modify the values in these columns, a dialog will appear to notify you that the change cannot be made (Figure 5. 3).

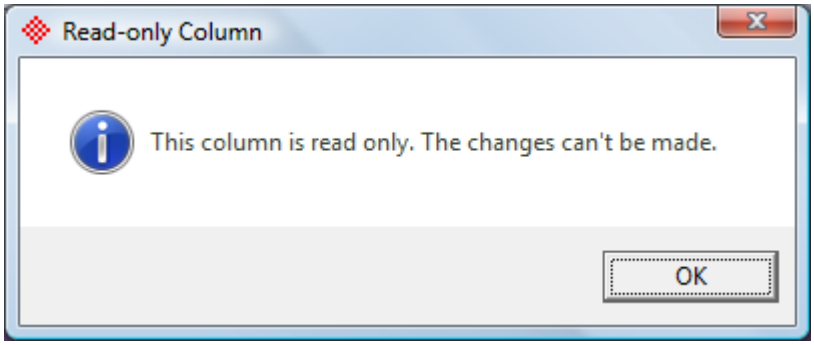


Figure 5. 3: Viewing the Read-only Column warning

Attribute Editing Limitations

The sample attributes need to be in English. It is recommended that you limit the maximum number of characters to 32 in a cell. Long or non-English attribute names will decrease the readability of visualizations.

Sample Attribute (column) Operation

To Select a Column

To select a column, click on the column header. The column will be highlighted.

To Add Column(s)

There are four ways to add columns to the table.

- End of the Table: Select the **Add Attribute…** button and choose a predefined attribute from the popup menu. The selected attribute will be added to the end of the table
- End of the Table: Select **Add Attribute…** > **Other…**. This will bring up the add column dialog (Figure 5. 4). Specify the desired column label and choose a column type using the radio buttons and select **OK**. The column will be added to the end of the table
- Before the Current Column: Right click on a column header and select **Insert Attribute**. This will bring up the *Add Column* dialog (Figure 5. 4). Specify the desired column label and choose a column type using the radio buttons and select **OK**. The column will be inserted before the column where the menu was invoked from

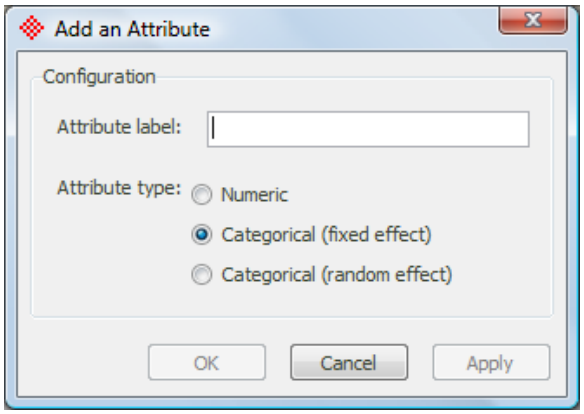


Figure 5. 4: Configuring the Add an Attribute dialog

- After the Current Column: Right click on a text column and select **Split Column.** This will bring up the Text to Column Splitter dialog (Figure 5. 5). Two options are provided to split the column – by delimiters or by fixed width. After specifying splitting parameters, select the **Update** button. A preview of the result will be shown at the bottom of the dialog. Specify column labels and properties using the entry box and dropdown menu. If column labels are not specified, default labels will be assigned automatically. To skip a column, choose **Skip** from the drop down menu under that column. Select **OK** when done. The resulting column(s) will be inserted after the column being split

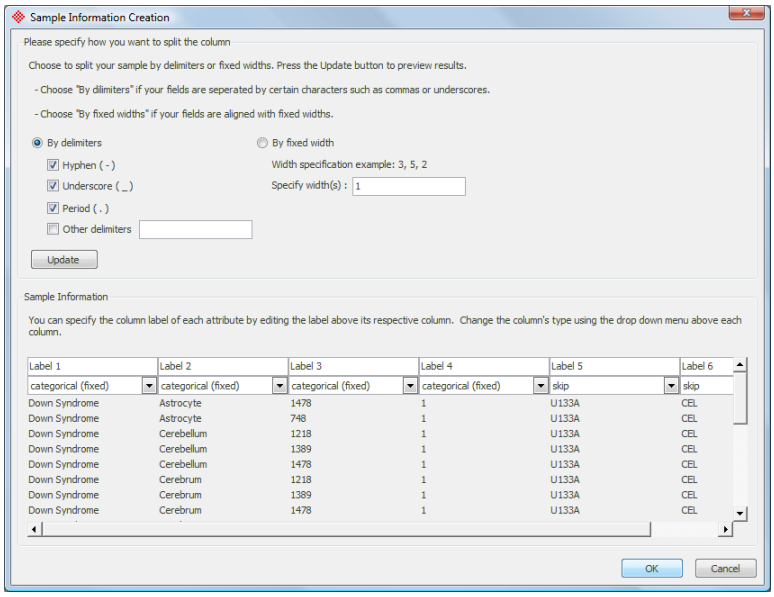

Sample Information Creation
Please specify how you want to split the column
Choose to split your sample by delimiters or fixed widths. Press the Update button to preview results.
- Choose "By dilimiters" if your fields are seperated by certain characters such as commas or underscores.
- Choose "By fixed widths" if your fields are aligned with fixed widths.
By delimiters
By fixed width
Hyphen (-)
Underscore (_)
Period (.)
Other delimiters
Width specification example: 3, 5, 2
Specify width(s) :
Update
Sample Information
You can specify the column label of each attribute by editing the label above its respective column. Change the column's type using the drop down menu above each column.
Label 1
Label 2
Label 3
Label 4
Label 5
Label 6
categorical (fixed)
skip
OK
Cancel

Figure 5. 5: Configuring the Sample Information Creation dialog

To Delete a Column

To delete a column, right click on the column header and select **Delete Attribute**.

To Edit Column Properties

Right clicking on a column header and selecting **Properties** will bring up the column properties dialog (Figure 5. 6). Specify the desired *attribute label* and *attribute type* and select **OK**. Note that the sample information editor won't allow two columns to have the same column label. If the new column label specified already exists in the table, the **OK** button will gray out.

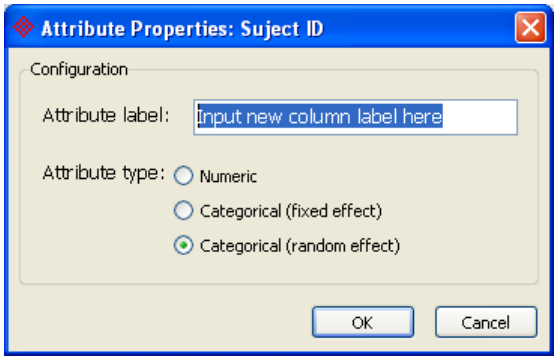


Figure 5. 6: Column properties dialog

To Reposition a Column

Repositioning a column can be done by drag and drop. Simply select the header of the column to be repositioned and start dragging, when the column is in the desired spot, release the mouse button.

To Sort Samples Based on One Column

To sort samples based on one column, right click on a column header and select **Sort Ascending** or **Sort Descending**.

To Sort Samples Based on Multiple Columns

To sort samples based on multiple columns, sort samples in the reverse order of the columns desired to be sorted. For example, if you want to sort samples first based on scan date then based on type, sort them based on type first and then scan date.

Sample (row) Operation

To Select Row(s)

To select rows, click on the row header. Multiple rows can be selected using the Control and /or Shift key. Row selection can also be done by clicking on a bar in the sample bar chart. All rows that fall in the category will be selected.

To Delete Row(s)

To delete row, select the rows to be deleted. After selecting the rows, right click on the row header and select **Delete Sample(s)**.

Edit Sample Attribute/Table Cell Operation

To Select Cells

To select a cell, click on the cell. Multiple cells can be selected by holding the mouse, dragging, and using the Control and /or Shift key. Note that only cells within a single column can be selected at the same time.

To Edit Multiple Cells

There are four ways to edit multiple cells.

- Select multiple cells and start typing. This will change the value of all the selected cells simultaneously
- Select one or more (consecutive) cells, select the bottom edge of the selected area and drag down
- Use copy, cut, and paste (see below)
- Right click on a bar in the sample bar chart and select *Change Value*. Edit the value in the *Edit Sample Attribute* dialog (Figure 5. 7) and select ***OK***. The value of the category will be changed to the new value

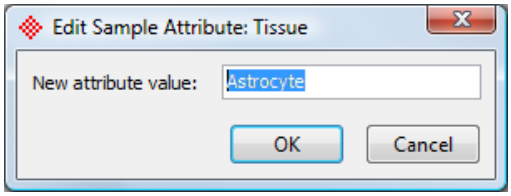


Figure 5. 7: Edit Sample Attribute Dialog

To Copy, Cut, and Paste

- To copy, select the cells and select **Copy** from right mouse menu, or press Control+c.
- To cut, select the cells and select **Cut** from right mouse menu, or press Control+x.
- To paste, select the cells and select **Paste** from right mouse menu, or press Control+v. Note that if multiple cells are selected, paste can only be performed if they are consecutive.

To Drag and Fill Data Automatically in Multiple Cells

Instead of entering sample information manually, you can use the auto fill feature provided by the Partek Express sample editor. The auto fill feature can identify certain data patterns and automatically fill in multiple cells. To start, select the cell(s) that contain the data that you want to fill into adjacent cells.

Note: these cells have to be in the same column and they must be in consecutive order.

- Place the mouse at the bottom edge of the selected cells. If auto fill is possible, the mouse cursor will turn into a down arrow. Hold the mouse down and drag down across the cells you want to fill.

The following patterns are supported for auto fill:

- Arithmetic series (integers, floating point numbers or integers with same prefix or suffix)
- Days of the week
- Months of the year

Undo and Redo Editing

You can undo and redo up to 100 actions per session in the Partek Express sample editor.

To Undo

To undo, select the **Undo** button or press Control+z.

To Redo

To redo, select the **Redo** button or press Control+y.

Note that certain actions cannot be undone, such as saving a study. If you can't undo an action, the **Undo** button will be grayed out.

Formatting the Table

To Best Fit Column(s)

There are two ways to best fit column width(s).

1. Select the **Fit Columns** button. This will best fit all columns in the table
2. Double click on the right edge of a column header. This will best fit the column left to it

To Manually Specify a Column Width

Right clicking on a column header and selecting **Column Width** will bring up the *Column Width* dialog (Figure 5. 8). Specify a positive integer pixel value as the column width and select **OK**.

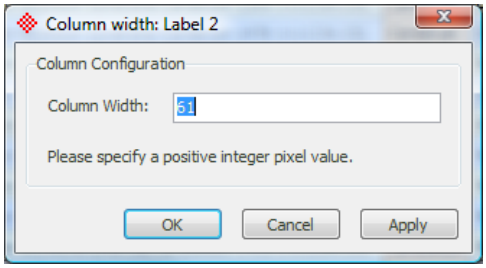


Figure 5. 8: Configuring the Column Width dialog

Partek® Express™: Data Import

Introduction

For most studies, Partek® Express™ will automatically download library/annotation files and import the data. This chapter provides more details on Partek Express' data import algorithm and on situations when automatic downloading cannot be achieved.

Import Algorithm

Import .CEL Files

When importing .CEL files, Partek Express uses RMA, which includes background correction, quantile normalization, and median polish summarization. For more information about RMA, refer to the following:

- Bolstad, B.M., Irizarry R. A., Astrand, M., & Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193.
- Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Terence, P., &Speed, T.P. (2003), Summaries of Affymetrix GeneChip probe level data Nucleic Acids Research 31(4):e15.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., & Speed, T.P. (2002) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics 4 (2):249-64.

Note: Median polish summarization may produce the same value for each sample when there are very few samples. For this reason, when there are <= 4 samples, Partek Express will use mean summarization.

Import .CHP Files

Intensities in .CHP files are already normalized and summarized. Partek Express directly reads the intensities without further processing the data. For more information, please refer to the Affymetrix software that was used to create the .CHP files.

Import Agilent Files

When importing Agilent two color files, Partek Express uses mean as the summarization method and does loess normalization to get the log ratio of two color data. When importing Agilent one color files, Partek Express does log base 2 transformation.

Import Illumina Files

Data from Illumina's BeadStudio software package can be exported in a custom Partek report file (.ppj file) for seamless importation of your Illumina data into Partek Express. When importing Illumina files, browse to select the .ppj file and select open.

Import NimbleGen Files

When importing NimbleGen files, Partek Express also uses RMA, which includes background correction, quantile normalization, and median polish summarization.

Library and Annotation Files

Affymetrix Library and Annotation Files

When importing .CEL files, Partek Express will automatically download the corresponding library files (Figure 6. 1).

Figure 6. 1: Downloading the Library File during data import

If the library files are stored on a local drive, select the abort button to manually specify the file location. As shown in Figure 6. 2, the *Specify File Location* dialog will appear.

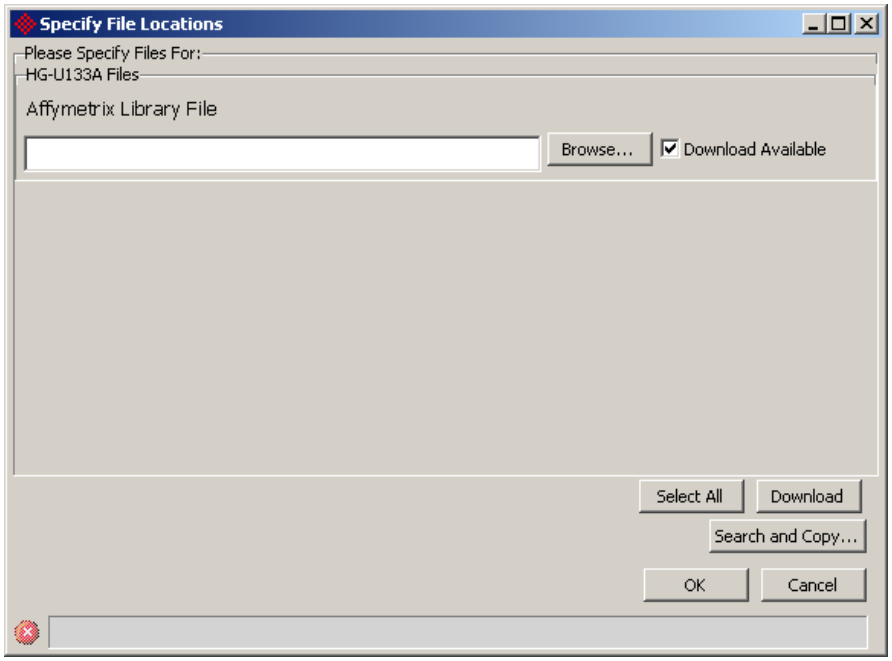


Figure 6. 2: Specifying the File Location

There are several options available to specify the library file:

1. Select the *Browse…* button to select the library file
2. Select the *Search and Copy…* button to let Partek Express find the correct library file by searching a directory and all its sub-directories
3. Select the *Download* button to continue the downloading process

After manually specifying the file location, select **OK**.

Agilent Annotation Files

When importing Agilent files, Partek Express will automatically generate an annotation file, which is extracted from the Agilent data files.

Illumina Annotation Files

Data from Illumina's BeadStudio software package can be exported into two files. One is a custom Partek report file (.ppj file) and the other is an annotation file (.annotation.txt). Place the annotation file into the directory that holds the .ppj file, so that Partek Express will be able to find the annotation file during import.

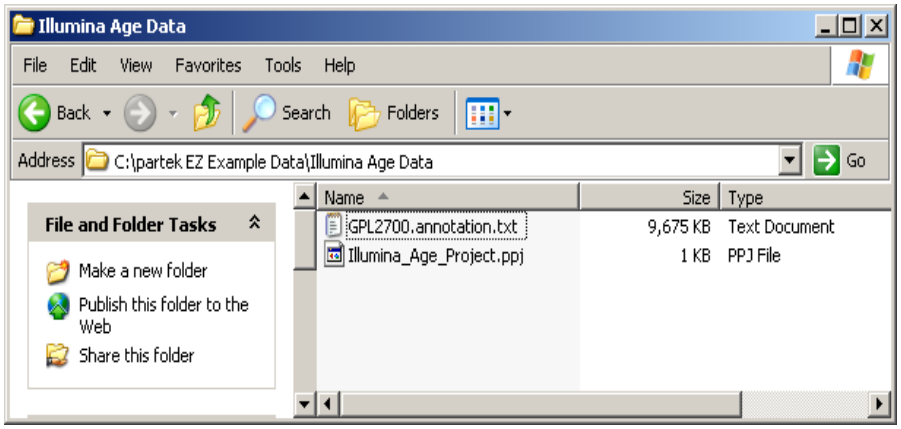


Figure 6. 3: Viewing the directory that holds the Illumina .ppj file and Annotation files

NimbleGen Annotation Files

The NimbleGen importer will invoke a window to specify a NimbleGen annotation file (.ngd file). After importing, this annotation file will be converted into a Partek recognized format (.annotation.txt).

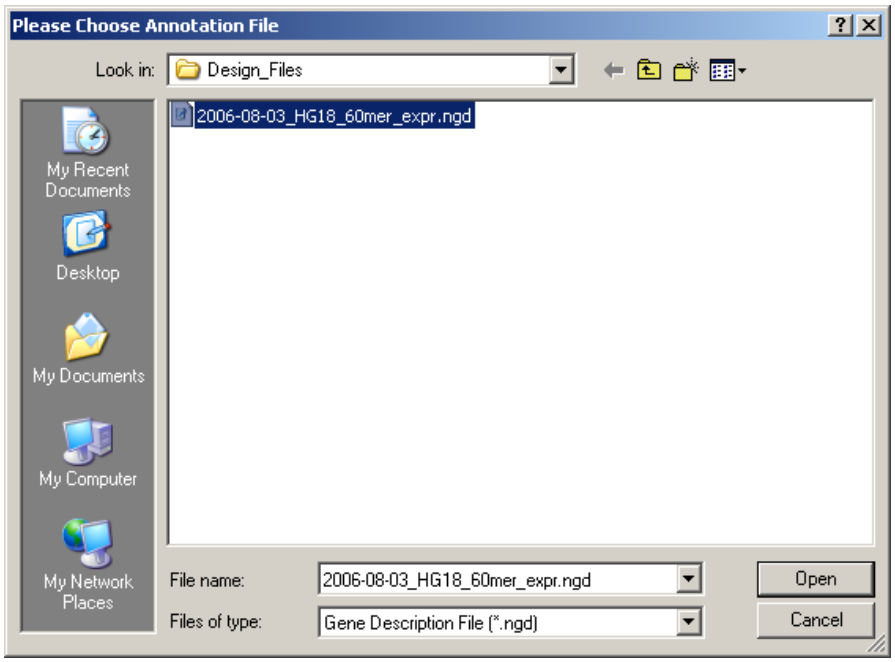


Figure 6. 4: Specifying the NimbleGen Annotation file

Partek® Express™: Quality Assessment

Introduction

The quality assessment procedures in Partek® Express™ are explained in this chapter. The procedures include generating QC metrics, invoking QC plots, and applying QC checking. The detailed algorithm of QC metrics also will be explained.

Quality Assessment Procedures

Generate QC Metrics and Graphs

QC metrics are generated from importing CEL files or CHP files; however, the CEL file importer invokes the QC metrics calculation during importing, but the CHP file importer just extracts QC metrics from the CHP files.

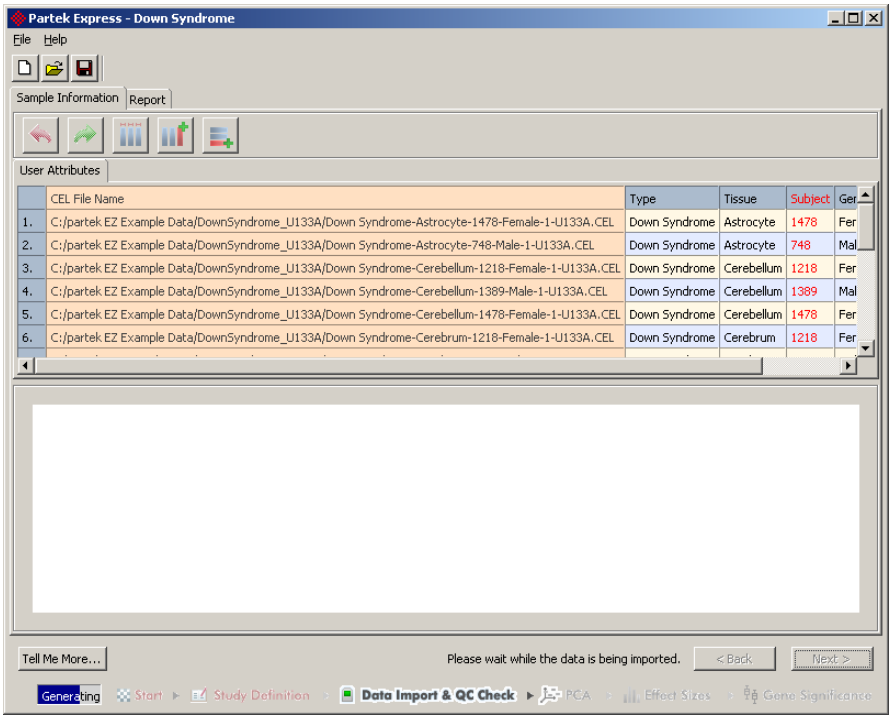

Partek Express - Down Syndrome
File Help
Sample Information Report
User Attributes
CEL File Name Type Tissue Subject
1. C:/partek EZ Example Data/DownSyndrome_U133A/Down Syndrome-Astrocyte-1478-Female-1-U133A.CEL Down Syndrome Astrocyte 1478
2. C:/partek EZ Example Data/DownSyndrome_U133A/Down Syndrome-Astrocyte-748-Male-1-U133A.CEL Down Syndrome Astrocyte 748
3. C:/partek EZ Example Data/DownSyndrome_U133A/Down Syndrome-Cerebellum-1218-Female-1-U133A.CEL Down Syndrome Cerebellum 1218
4. C:/partek EZ Example Data/DownSyndrome_U133A/Down Syndrome-Cerebellum-1389-Male-1-U133A.CEL Down Syndrome Cerebellum 1389
5. C:/partek EZ Example Data/DownSyndrome_U133A/Down Syndrome-Cerebellum-1478-Female-1-U133A.CEL Down Syndrome Cerebellum 1478
6. C:/partek EZ Example Data/DownSyndrome_U133A/Down Syndrome-Cerebrum-1218-Female-1-U133A.CEL Down Syndrome Cerebrum 1218
Tell Me More...
Please wait while the data is being imported.
< Back
Next >
Generating
Start
Study Definition
Data Import & QC Check
PCA
Effect Sizes
Gene Significance

Figure 7. 1: Generating QC metrics during importing

Upon finishing the import, the quality assessment results will be shown in the *QC Metrics* tab. The *QC Metrics* tab provides quality control information from the control and experimental probes on the Affymetrix chips to provide confidence in the quality of the microarray data or to identify samples that do not pass the QC metrics. The QC metrics result can be viewed either in line graph format or in

spreadsheet format by selecting the corresponding radio buttons at the top of *QC Metrics* tab.

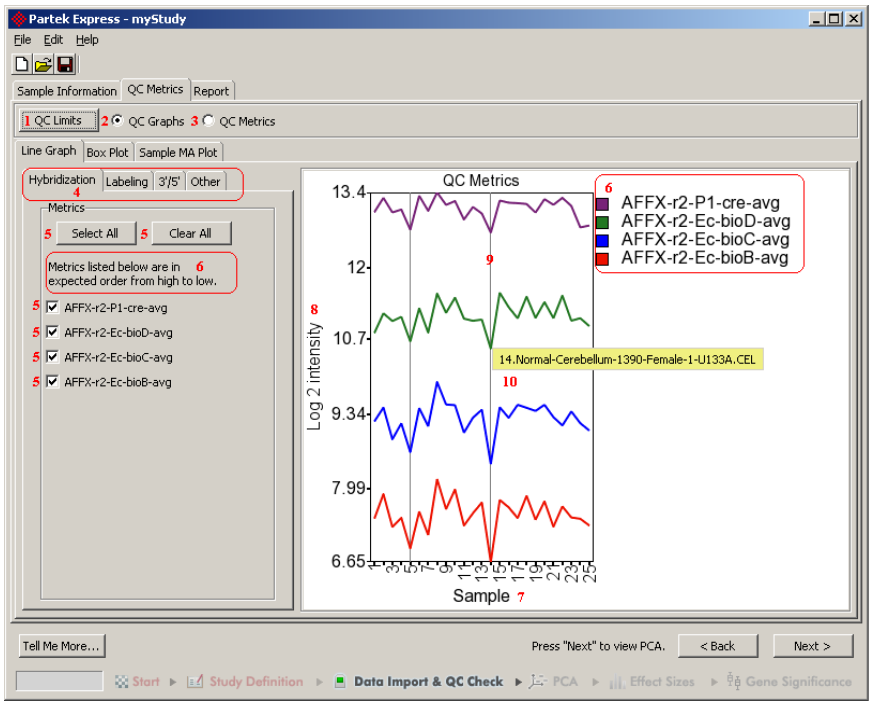
The main screen parts are explained below and pictured in Figure 7. 2.

1. The *QC Limits* button; selecting this will perform QC limits checking
2. The *QC Graphs* radio button; selecting this will allow you to view QC graphs including line graph, sample box plot and sample MA plot by selecting the corresponding tab
3. The *QC Metrics* radio button; selecting this will allow you to view the QC metrics table
4. QC metrics grouping categories, which includes *Hybridization, Labeling, 3'/5'* and *Other*. By default, the *Hybridization* tab is selected; to view the other categories, select the corresponding tab
5. The *Select All* and *Clear All* buttons; selecting will either check or uncheck QC metrics
6. The *Hybridization and Labeling* controls' metrics are listed in the expected order from high to low. In this example, AFFX-rs-P1-cre-avg should be higher than AFFX-r2-Ec-bioD-avg, which should be higher than AFFX-r2-Ec-bioC-avg, which should be higher than AFFX-r2-Ec-bioB-avg. For the controls on *3'/5'* and *Other* tab, their QC metrics will not be listed in any particular order.
7. Samples are on X axis. In this example, there are 25 samples. Their order is the same as on the *Sample Information* tab.
8. For *Hybridization and Labeling* controls, their metrics are in *Log 2 intensity* scale. For *3'/5'* controls, their metrics are ratios*. For metrics under the *Other* tab, the Y axis may be *Intensity* or *Value.*
9. The *sample selection lines;* <click> or <control-click> to select one or more samples. The sample selection is synchronized with the *Sample Information* tab and sample selection boxes in the box plot
10. The mouse-over view in the Line Graph; dragging the mouse over points in the graph will show the .CEL file name of a sample
11. The *Log Expression Signal* radio button; selecting this will allow you to view a box plot on the log expression signal. This box plot is generated from the probe set signal values that have been normalized and summarized
12. The *Log Probe Cell Intensity* radio button; selecting this will allow you to view a box plot on the log probe cell intensity. This box plot is generated from the probe cell intensity values prior to normalization and summarization. This box plot is only available for CEL file importer but not CHP file importer, because the data from CHP file are already normalized and summarized
13. The *Relative Log Expression Signal* radio button; selecting this will allow you to view a box plot on the relative log expression signal. This box plot is a summarization of the RLE values which are obtained by calculating

the log base 2 difference between the probeset signal estimate and the median of probeset signal estimates across all of the arrays

14. *The sample selection boxes*; <click> or <control-click> to select one or more samples. The sample selection is synchronized with the Sample *Information* tab and sample selection lines in the line graph
15. The mouse-over view in the Box Plot; dragging the mouse over points in the graph will show the five-number summaries of each sample. The five-number summaries are the smallest observation, 25th quantile (Q1), median, 75th quantile (Q3), and the largest observation
16. Sample MA Plot; selecting sample 1 and sample 2 will invoke a sample MA plot, which compares the intensities between these two samples (rows). Each point represents one column in the spreadsheet. For the two selected samples, the average is on the X axis, and the difference on the Y axis. The points are expected to be centered along Y=0 for all values of X

* When importing .CEL files, Partek Express transforms the intensity with log base 2. Suppose the log 2 intensity of 3' control is A3 and the log 2 intensity of 5' control is A5, Partek Express first does base 2 anti-log A3 and A5 to a3 and a5 then calculates the ratio as a3/a5. When importing .CHP files, Partek Express directly reads the QC metrics from the header of .CHP files. Please note some .CHP files may have calculated 3'/5' ratio from the logged data. For example, the intensity of normalized 3' control is A3 and 5' control is A5. They have already been logged. Some .CHP files may directly give the ratio as A3/A5.



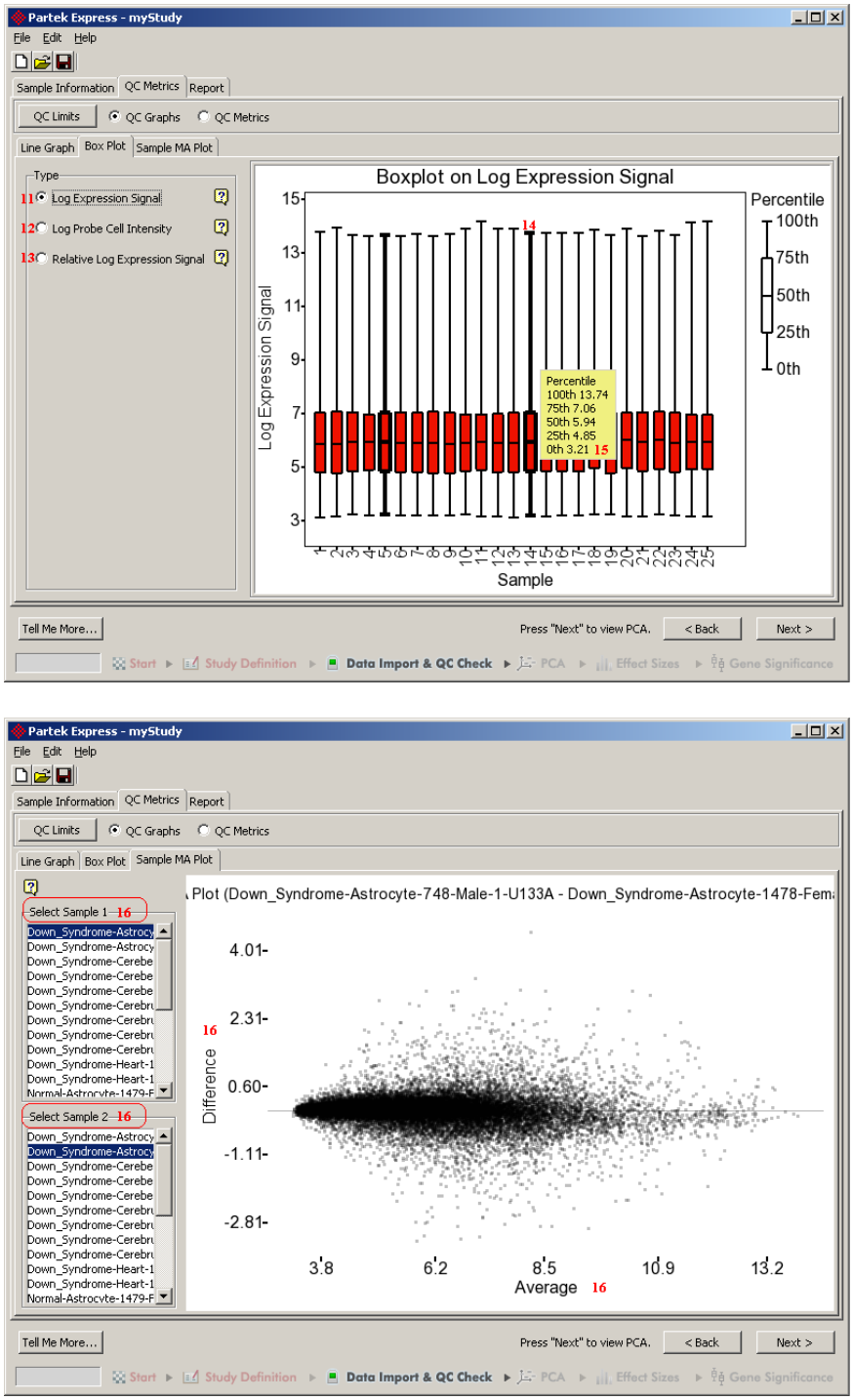


Figure 7. 2: Viewing the QC metrics visualizations

Apply QC Checking

Introduction

When the QC metrics data is generated, the QC data is automatically tested against several predefined criteria. If any of the QC data fail any of the criteria, the failing QC metrics will be highlighted in the QC metrics spreadsheet at which point a determination must be made to either continue the analysis, omit the samples that failed the QC criteria, or to rerun the failed samples to generate new data that passes the QC criteria.

The main screen parts of the *QC Metrics* table are explained below and pictured in Figure 7. 3.

1. The *QC Metrics* radio button; selecting this will allow you to view the *QC Metrics* table
2. Metrics that didn't pass the *QC Limits* checking will be highlighted
3. The sample selection is synchronized with the *Sample Information* tab

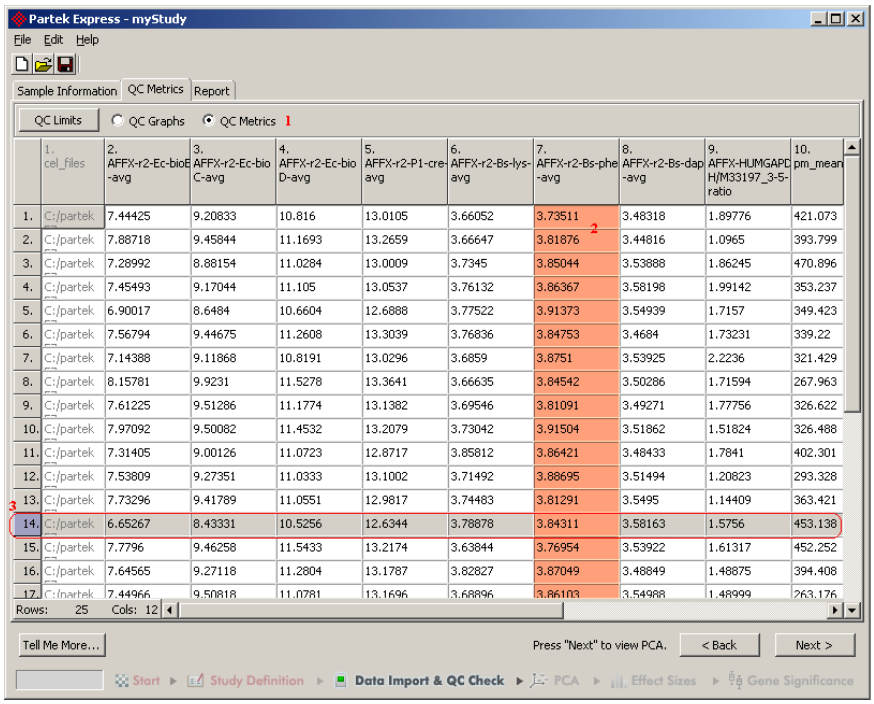


Figure 7. 3: Showing QC metrics table in Partek Express

Configuring the QC Check Dialog

Open the *Apply QC Checking* dialog by selecting the **QC Limits** button.

Pre-defined Thresholds

There are eight pre-defined thresholds within the *Apply QC Checking* dialog. These pre-defined thresholds are grouped according to their probe set properties, which

include *Pre-defined Hybridization, Pre-defined Labeling, Pre-defined 3'/5',* and *Custom.*

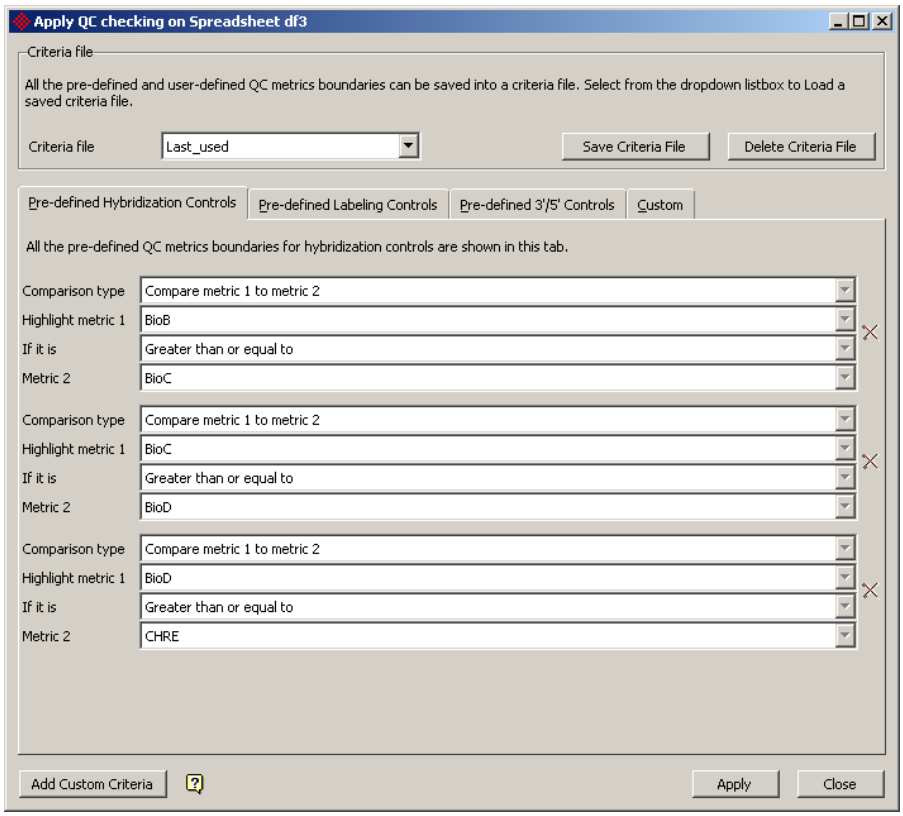


Figure 7. 4: Configuring the Applying QC Checking dialog

User Specified Thresholds

Selecting the **Add Custom Criteria** button at the bottom of the *Apply QC Checking* dialog will add user specified thresholds. The user specified thresholds will be shown on the *Custom* tab.

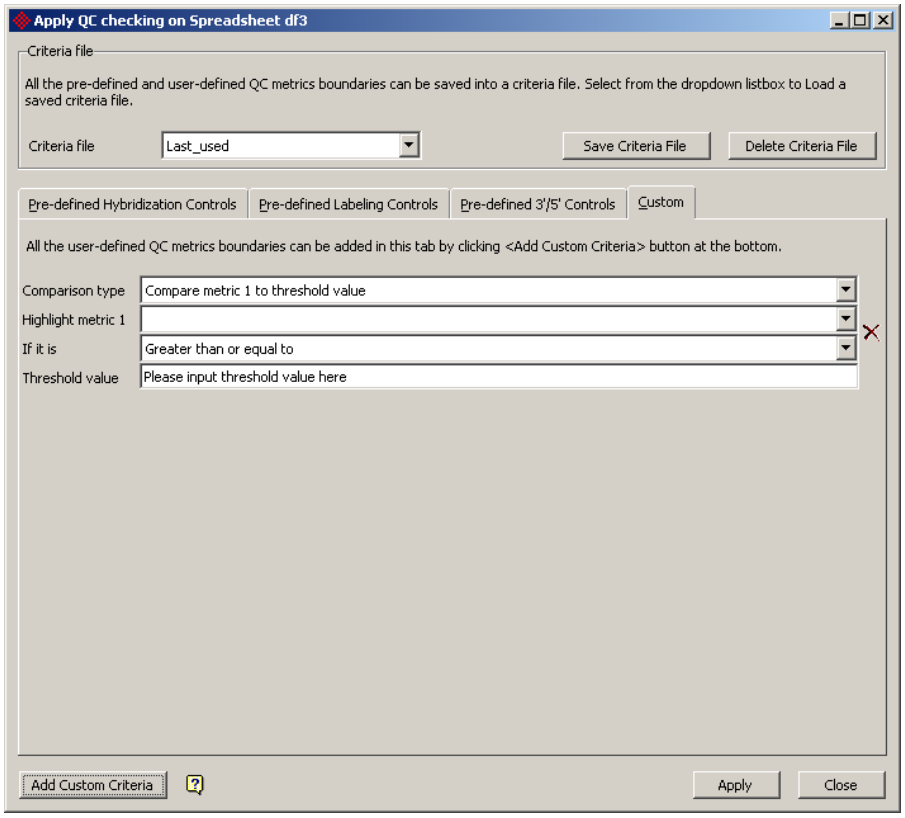

Apply QC checking on Spreadsheet df3
Criteria file
All the pre-defined and user-defined QC metrics boundaries can be saved into a criteria file. Select from the dropdown listbox to Load a saved criteria file.
Criteria file
Last_used
Save Criteria File
Delete Criteria File
Pre-defined Hybridization Controls
Pre-defined Labeling Controls
Pre-defined 3'/5' Controls
Custom
All the user-defined QC metrics boundaries can be added in this tab by clicking <Add Custom Criteria> button at the bottom.
Comparison type
Compare metric 1 to threshold value
Highlight metric 1
If it is
Greater than or equal to
Threshold value
Please input threshold value here
Add Custom Criteria
Apply
Close

Figure 7. 5: Adding user specified thresholds

The options in the *Custom* tab include the following:

1. Comparison type:
    1. Compare metric 1 to metric 2
    2. Compare metric 1 to threshold value
    3. Compare metric 1 to average of metric 1 across arrays by standard deviation. The in boundary range for metric 1 is (mean - range multiplier * standard deviation, mean + range multiplier * standard deviation)

2. Comparison operator (if it is):
    1. Less than
    2. Less than or equal to
    3. Greater than
    4. Greater than or equal to
    5. Equals
    6. Not equals

Select the metric for Metric 2 or input the threshold value or the range multiplier. The option to be used here is decided by the comparison type selected above. ✕ will delete the user specified threshold. Pre-defined thresholds cannot be deleted.

Save Thresholds

All the user specified thresholds combined with the pre-defined thresholds can be saved as a criteria file by selecting the **Save Criteria File** button. The entry box for the criteria file name is located on the top of the dialog. The *Default* thresholds are created by *Partek Express* and cannot be overwritten. The thresholds that were applied last time will be automatically saved as a criterion file named *Last_used*, which is also not able to be modified.

Delete Thresholds

Delete Criteria File can delete any criteria file you save. The criteria file *Default* and *Last_used* are created by *Partek Express* and cannot be deleted.

Load Thresholds

The previously saved thresholds can be loaded into *Partek Express* by selecting the criteria file name from the dropdown list box, which is also used as an entry box for saving thresholds. All the criteria files that have been saved will be shown in the dropdown list box. *Last_used* criteria will always be loaded by default.

Running the QC checking

Apply will check if the QC metrics results are within the boundaries defined by those thresholds. Any result that is out of the boundary will be colored on the QC metrics table (Figure 2).

Cancel will close the QC Check dialog without doing any checking.

Quality Assessment Metrics

Hybridization Metrics

Four exogeneous (*E. coli* derived) pre-labeled molecules are spiked into the hybridization cocktail before hybridization, but after sample labeling. The spikes test to ensure that hybridization correctly occurred on the array. These molecules are spiked in at increasing concentrations: BioB < BioC < BioD < Cre. A graph of the values is automatically created and displayed in the *Hybridization* tab within the *QC Metrics* section. Make sure to ensure that each of the spikes has the correct relative abundance in the samples as displayed in Figure 7. 6.

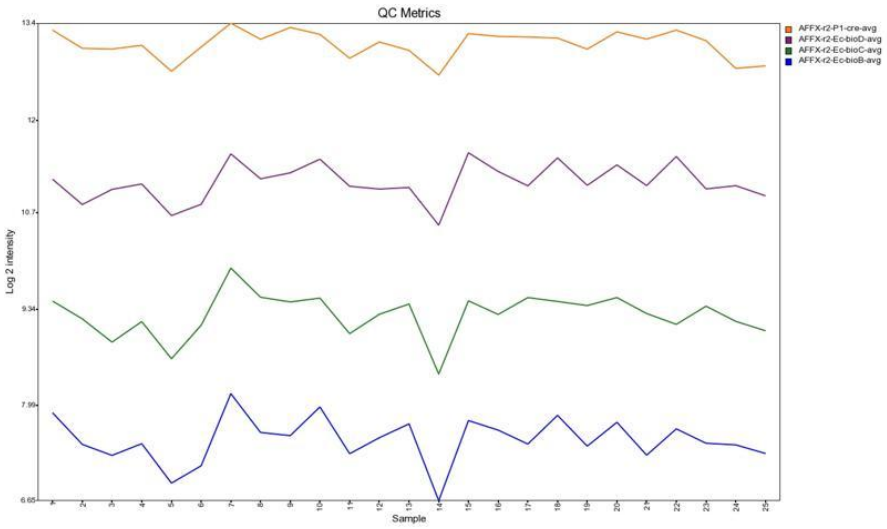


Figure 7. 6: Viewing the line graph of Hybridization Spikes. In each sample, the four hyb spikes have increasing concentrations from BioB as the lowest to Cre as the highest.

Labeling Metrics

Up to five unlabeled polyA control spikes are available to spike into the samples to control for the sample labeling reaction. The spikes are inserted into the sample prior to labeling and their resulting detection is dependent on the labeling reaction that labels the biological sample. They are derived from B. subtilis, and are typically spiked in at increasing concentrations of Lys < Phe < Thr < Dap. Make sure to confirm that these spikes were used in the samples and confirm that the correct concentrations were used. This Down's syndrome experiment was run before these spikes were commercialized and they show a different intensity pattern. The graph of these spikes (Figure 7. 7) is displayed in Partek Express in the *QC Metrics* section under the *Labeling* tab. Partek Express only extracts the Dap, Phe, and Lys spikes.

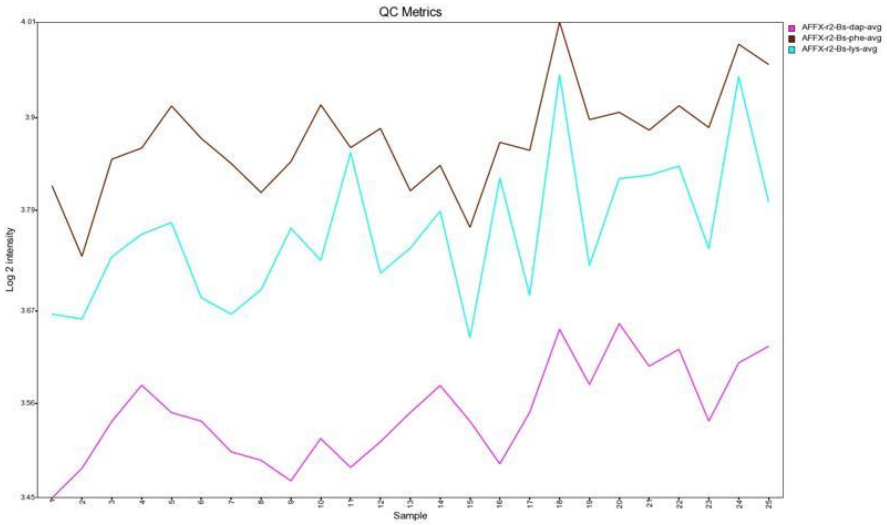


Figure 7. 7: Labeling spikes of Dap, Phe, and Lys. This experiment shows DAP < LYS < PHE

3' / 5' Ratio Metrics

Partek Express will calculate and plot the 3' / 5' ratio of GAPDH. It is displayed under the *QC Metrics* section in the *3'/5'* tab. GAPDH has separate probe sets at the 3' and 5' end of the gene. In high-quality samples, reverse transcriptase should process from the 3' through towards the 5' end. The 3' / 5' ratio compares the abundance of the signal at the 3' end over the abundance at the 5' end. A ratio of 3 or less is considered acceptable.

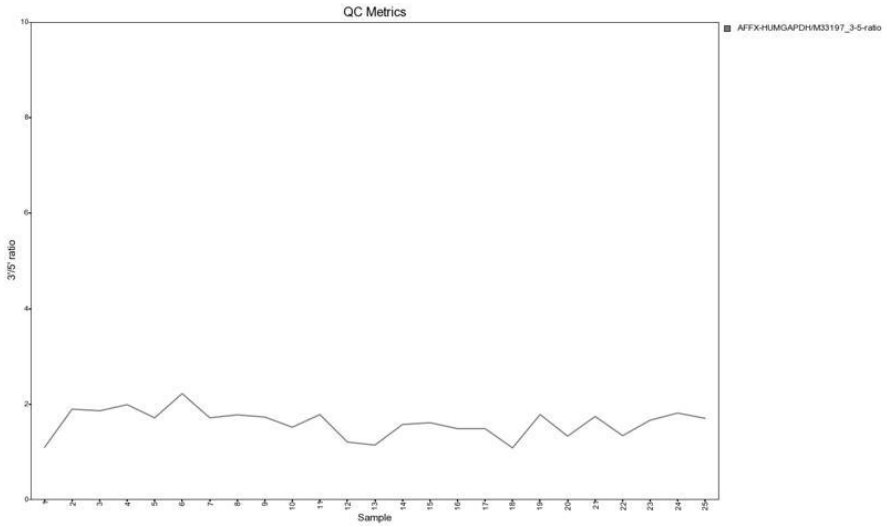


Figure 7. 8: Viewing the 3' / 5' Ratio for Human GAPDH across all samples in the experiment; all values are less than

Other Quality Control Metrics

Three additional quality control metrics are displayed in the *Other* tab within the *QC Metrics* section: *PM Mean, Mad_Residual Mean*, and *RLE Mean*. For more information on these values consult the Quick Reference Card from Affymetrix entitled, "QC Metrics for Exon and Gene Design Expression Arrays".

PM Mean

PM Mean is the mean raw probe intensity from a sample. It is a measure of how bright or dim an array is. Samples within an experiment should have roughly similar PM Means. There are not any default criteria regarding PM Mean. Samples should be scanned for "outlier" values as determined by the user through visual inspection.

MAD Residual Mean

MAD Residual Mean is a bit of a complex measurement. It is the mean across all probe sets of the Median Absolute Deviation (MAD) of the residuals between the predicted and actual probe values. During signal estimation, a model is created

based on the trends for each probe across the whole experiment. This model can be used to "predict" how a probe will respond. The residual is the difference between the predicted and actual values. When examined at a sample level (across all probe sets) the MAD Residual Mean value is a measure of how well the individual sample fits the model for the experiment. Samples with higher values fit less well.

RLE Mean

RLE Mean is the mean of the absolute relative log expression (RLE) across all probe sets on each array. Consult Chapter 6 of the Partek Express User Manual for more information on its calculation. RLE Mean compares the signal each probe set (gene) in a sample compared to the median gene-level signal value across the experiment (all samples). If a sample has a high RLE Mean that implies that that sample isn't quite as similar to all of the samples. High RLE Mean values will flag outliers. Affymetrix states that RLE Mean values across a diverse tissue panel range from 0.27 to 0.61, while values across an experiment of only technical replicates range of 0.1 to 0.23. Remember that if you have a collection of diverse samples in the experiment the RLE Mean values will be higher than if the samples were very similar.

References

Affymetrix White Paper: Quality Assessment of Exon and Gene Arrays. Revision 1.1, published April 06, 2007. http://www.affymetrix.com/support/technical/whitepapers/exon_gene_arrays_qa_whitepaper.pdf.

Affymetrix Quick Reference Card, "QC Metrics for Exon and Gene Design Expression Arrays" http://www.affymetrix.com/Auth/support/downloads/quick_reference_cards/qc_metrics_exon_gene_qrc.pdf

Partek® Express™: Principal Components Analysis

Introduction

This section describes the principal components analysis (PCA) plot and its function within Partek® Express™.

Invoking the PCA Plot

- Select the **Next** button after completing the *Data Import & QC Check* step. The PCA plot will be brought up and shown in another tab named *PCA* (Figure 8. 1)

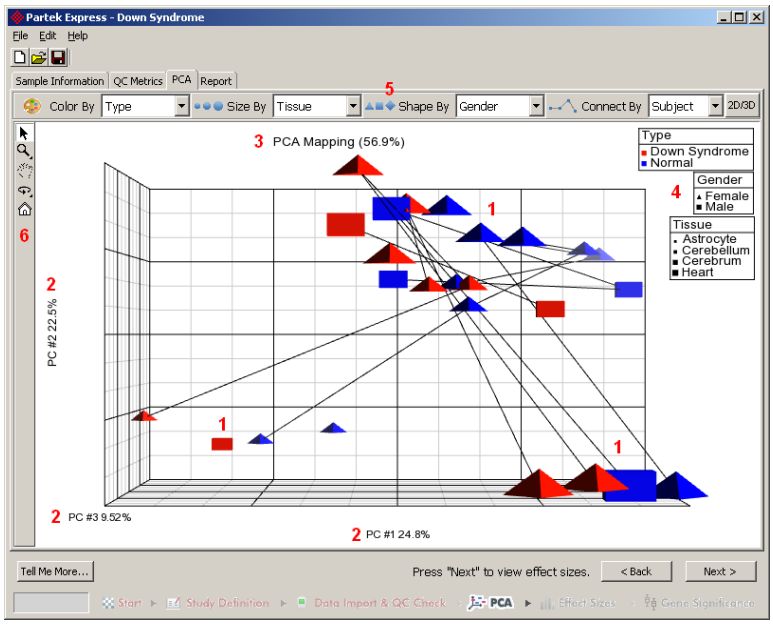


Figure 8. 1: Viewing the PCA Plot

The main screen parts are:

1. The *Shape by* elements; these can be shaped as spheres, cubes, and/or pyramids, etc. They represent individual samples. For example, in this study there are 25 samples, represented by the 25 elements in the PCA plot. The grouping of, and the relative distance between, those objects visually reveal the relation of those samples
2. The X, Y, and Z axes are PC #1, PC #2, and PC #3, respectively. The percentages shown in the axis labels display the amount of data variation accounted for by the respective PC. Please refer to the *Details About PCA* below for more information. Since the PC's are ordered from greatest to

smallest amount of explained variation, PC #1 will be greater or equal to PC #2, which is greater than PC #3

3. The percentage shown in the title is the sum of the percentage on those axes. It cannot be greater than 100%
4. The *Legends*; these describe the use of symbols, colors, etc., used in the plot
5. The *Tools* bar; selecting options within the *Tools* bar will change the PCA plot style. See *Configuring the PCA Plot* below for more details
6. The *Mode* bar; options located in the *Mode* bar will control the PCA plot. See *Using the Viewer Modes in the PCA Plot* below for more details

Viewing the PCA Plot

PCA is an excellent method for visualizing high dimensional data by reducing the variation across all of the many thousands of probes being interrogated on the chip into a two or three dimensional representation. In a PCA plot, each point represents a sample (microarray) and corresponds to a row in the *Sample Information* tab. The positions of the dots are relative to each other. Those dots that are closer to each other represent samples in which the transcriptome measurements over the whole chip are similar. Those dots that are further away from each other represent samples in which the transcriptome measurements over the whole chip are more dissimilar. Samples that have similar overall gene expression levels will group together into clusters. Identifying separate clusters in a PCA provides valuable information, such as, which of the phenotypic variables are driving the major sources of variation within the experiment.

One example would be if an experiment only had one factor, treated and untreated. Assuming that all the samples in the data set are the same except for this one factor, it is possible to quickly identify if the treatment had a significant effect on the overall gene expression. If all of the samples clustered together into one group with the two colors mixed equally among the cluster, then there is no distinctive difference between the gene expressions over the samples based upon treatment. However, if the samples cluster into two distinct groups, one cluster containing only treated samples and the other cluster containing only untreated samples, then there is a difference in the gene expression profiles between the treated and untreated samples.

Details About PCA

PCA is an exploratory technique that is used to describe the structure of high dimensional data by reducing its dimensionality (Jolliffe, 1986). It is a linear transformation that converts *n* original variables into *n* new variables ("PC's"), which have three important properties:

- The new variables (PC's) are ordered by the amount of variance explained
- The new variables (PC's) are uncorrelated

- The new variables (PC's) explain all variation in the data

PCA is a *Principal Axis Rotation* of the original variables that preserves the variation in the data. Therefore, the total variance of the original variables is equal to the total variance of the principal components. The *eigenvectors* and *eigenvalues* define the rotation and variation and are described as follows:

- The *eigenvalues* are the variances of the principal components
- The *eigenvectors* are the direction cosines of the new axes (PCs) relative to the old (original variables), thus they define the rotations of the original axes

The method of PCA dates back to Harold Hotelling's 1933 paper "Analysis of a complex of statistical variables into principal components".

Configuring the PCA Plot

Style

Color, size, shape and connecting lines can be configured on the toolbar on the top of the PCA viewer.

Color By

By default, the color of the points is determined by the first categorical attribute in the sample information spreadsheet. If the sample information spreadsheet does not have a categorical variable, then the points will be colored by the first attribute. By choosing from the *Color* drop down menu (Figure 8. 2), the points can be colored by any column or by all the same color. If a categorical attribute is selected, each category will be assigned a distinct color. If a numeric attribute is selected, the color assigned to points will be based on a continuous color palette.

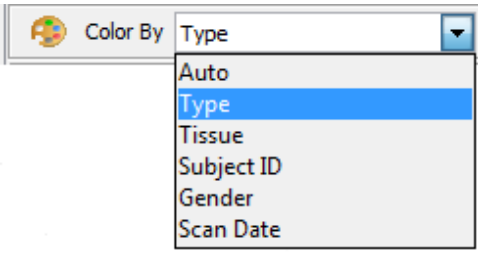


Figure 8. 2: Viewing the options in the Color By Configuration Menu

Size By

By default, all points in the plot are of the same size. Sizes of points can be configured by choosing from the *Size* drop down menu (Figure 8. 3). If size is set to "**Auto**", the size of the points will be based on the number of points in the plot. The size of a point can also be configured to correspond to the value in a specific

column in the sample information spreadsheet. When sizing by a categorical attribute, each category will be assigned a distinct size. When sizing by a numeric column, the size will be based on the order of points, and the legend lists the minimum, middle, and maximum values.

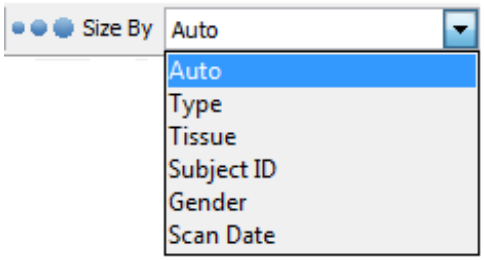


Figure 8. 3: Viewing the options in the Size By Configuration Menu

Shape By

By default, all points in the plot will be the same (point). Shapes of points can be configured by choosing from the *Shape* drop down menu (Figure 8. 4). When an attribute is selected, the shape of a point is determined by the attribute value of the corresponding sample.

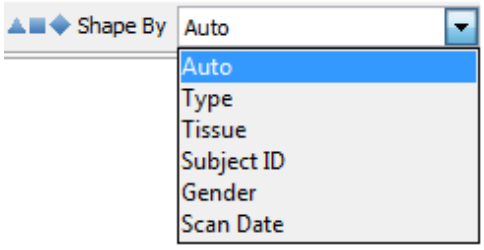


Figure 8. 4: Viewing the options in the Shape By Configuration Menu

There are five possible shapes to use. They are *sphere*, *tetrahedron*, *cube*, *octahedron*, and *icosahedron*. If the specified attribute is categorical and has 4 or fewer categories, each category will be assigned a distinct shape. If more than 5 categories are present in a categorical column, it will result in using the same shape for more than one category. If the specified attribute is numeric, the range of the values in that column is divided into 5 groups of equal range. The points will be shaped according to the group into which they fall.

Connect By

A line can be drawn among points that have the same value for the specified attribute. This is useful when looking at samples from the same subject (Figure 8. 5).

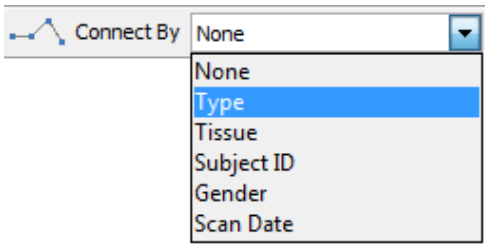


Figure 8. 5: Viewing the Connect By Configuration Menu

Dimension

When the dimension is set as *3D*, all X, Y, and Z axes will contain one PC from the dataset. If the dimension is set to be *2D*, only X and Y will be drawn and the viewer perspective will be turned off. The default setting for PCA dimension is *3D*.

Using the Viewer Modes in the PCA Plot

This section will explain how to use the *Mode Tool Bar* in the viewer (Figure 8. 6). The Mode Toolbar is the vertical toolbar at the left side of the viewer.

Figure 8. 6: Viewing the Mode Toolbar

Changing Modes

Most of the icons in the vertical mode toolbar have multiple options that can be accessed by clicking and briefly holding down the left mouse button on the mode icon, upon doing that, a mode option menu will pop-up to the right. To select an option from the menu, drag the mouse cursor over to the desired mode option and then release the mouse button.

Selection Mode

To invoke selection mode, click on the *Selection Mode* icon on the Mode Toolbar. When in selection mode, the following operations can be performed:

- Select an individual item – click the left mouse button
- Create a bounding box - hold down the left mouse button and drag the mouse creates a box to select items inside the box
- Add the item under the mouse cursor (or in the box) to the list of selections - **<Ctrl>** + left click

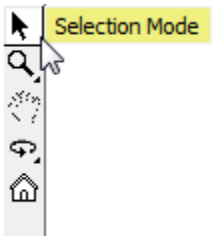


Figure 8. 7: Viewing the Selection Mode icon on the Mode Toolbar

Zoom Mode

To invoke *Zoom Mode*, click on the *Zoom Mode* icon on the Mode Toolbar. When in *Zoom Mode*, left click to incrementally zoom in, <**Ctrl**>-click to zoom out.

Figure 8. 8: Viewing the Zoom Mode icon on the Mode Toolbar

Pan Mode

This icon is enabled only when the data is zoomed in on. Hold down the left mouse button while dragging the mouse to interactively move the data (pan).

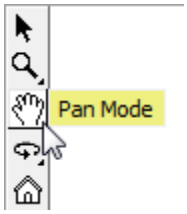


Figure 8. 9: Viewing the Pan Mode icon on the Mode Toolbar

Rotate Mode

There are two rotation modes: *Manual Rotation Mode* (one circle) and *Continuous Rotation Mode* (two circles, one is on top of another).

Figure 8. 10: Viewing the Rotation Mode icon on the Mode Toolbar

Figure 8. 11: Viewing the Manual Rotation Mode and the Continuous Rotation Mode

Manual Rotation Mode

The *Manual Rotation Mode* has the same functionality of the middle-mouse button. Hold down the left mouse button while dragging the mouse to interactively rotate the view.

Continuous Rotation Mode

Click the left mouse button to start and stop rotation. Selecting any other mode also stops continuous rotation.

Common in All Modes

Mouse over

Place the mouse cursor over a data item (without clicking) to see information about the item.

Reset

The <**Home**> key resets *Rotation*, *Zoom* and *Pan* back to their default values. The same can be achieved by clicking the *Reset* icon on the mode toolbar (Figure 8. 12).

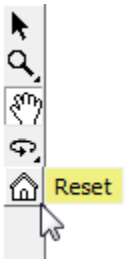


Figure 8. 12: Viewing the Reset Mode icon on the Mode Toolbar

References

Hotelling, H. “Analysis of a complex of statistical variables into principal components”. J. Educ. Psych 1933, 26: 417-441.

Jolliffe, I.T. *Principal Component Analysis*, Springer-Verlag, New York, 1986.

Partek® Express™: Effect Sizes

Introduction

Effect size provides information on the importance of each experimental factor to the transcriptome. Partek® Express™ uses Analysis of Variance (ANOVA) to test for the difference in means of a response variable between different groups.

Configuring the ANOVA Dialog

Before estimating effect sizes, sample information has to be created first. Please refer to the *Edit Sample Information* chapter to get more information about how to create sample information.

The Partek Express workflow will introduce you to the *Effect Sizes* step automatically by continually selecting **Next** > on the main window. The *Effect Sizes* step is after this sequence: **Start > Study Definition > Data Import & QC Check > PCA**.

Selecting Factor(s) of Interest

Main Effects

All the factors will be shown on the left panel of *Estimate Effect Sizes* dialog (Figure 9. 1), however only categorical (fixed) factors or numeric factors can be selected as the effect of interest. At most, two factors can be selected as the effect of interest. Drag and drop the factor from the left panel to the top or the middle panel on the right to set the effects of interest. The second effect of interest is optional. Partek Express will automatically detect whether a factor is estimable or not based on the current ANOVA model configuration. *Non-Estimable* will be inserted before the factor name once the factor is detected as non-estimable (Figure 9. 2).

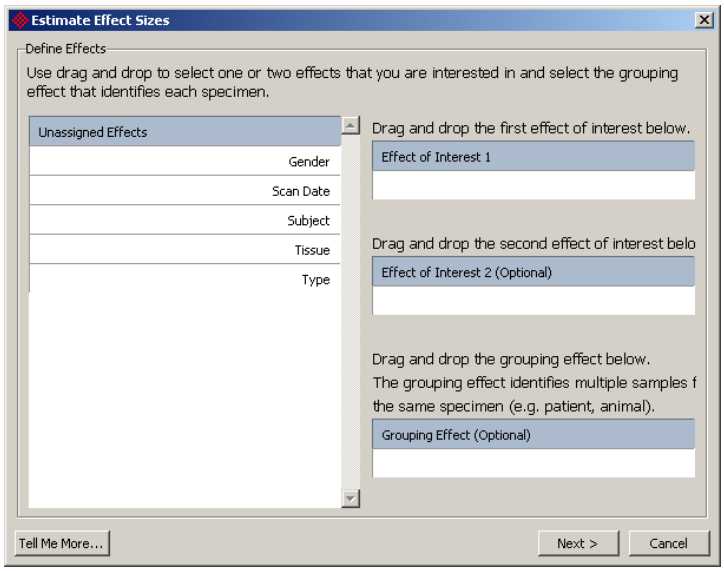

Estimate Effect Sizes
Define Effects
Use drag and drop to select one or two effects that you are interested in and select the grouping effect that identifies each specimen.
Unassigned Effects
Gender
Scan Date
Subject
Tissue
Type
Drag and drop the first effect of interest below.
Effect of Interest 1
Drag and drop the second effect of interest belo
Effect of Interest 2 (Optional)
Drag and drop the grouping effect below.
The grouping effect identifies multiple samples f
the same specimen (e.g. patient, animal).
Grouping Effect (Optional)
Tell Me More...
Next >
Cancel

Figure 9. 1: Estimate Effect Sizes, Dialog 1

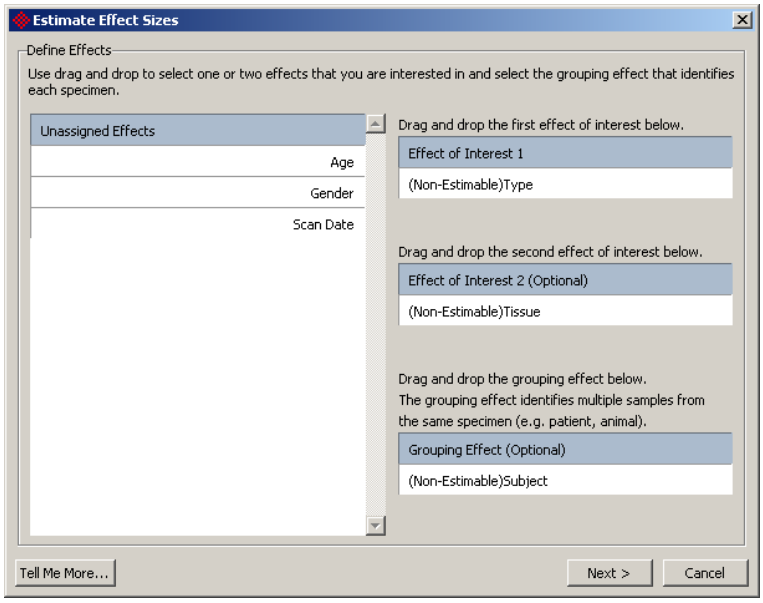

Estimate Effect Sizes
Define Effects
Use drag and drop to select one or two effects that you are interested in and select the grouping effect that identifies each specimen.
Unassigned Effects
Age
Gender
Scan Date
Drag and drop the first effect of interest below.
Effect of Interest 1
(Non-Estimable)Type
Drag and drop the second effect of interest below.
Effect of Interest 2 (Optional)
(Non-Estimable)Tissue
Drag and drop the grouping effect below.
The grouping effect identifies multiple samples from the same specimen (e.g. patient, animal).
Grouping Effect (Optional)
(Non-Estimable)Subject
Tell Me More...
Next >
Cancel

Figure 9. 2: Estimate Effect Sizes, Non-Estimable effects

Interactions

An interaction of the main effects will be automatically added to the ANOVA model if more than one main effect is selected. An interaction is the variation among the differences between means for the different levels of a factor over different levels of the other factor. However if the interaction is detected as non-estimable or it will cause some other main effect to be non-estimable, it will not be added to the ANOVA model.

Selecting the Grouping Factor

If there are multiple samples from the same specimen, the factor that holds the specimen is typically specified as the grouping factor. Drag and drop the factor from the left panel to bottom panel on the right to set the grouping effect.

Creating Comparisons

- Selecting **Next >** will bring up the *Create Comparison* dialog (Figure 9. 3), which allows you to perform a linear contrast between two specific groups within the context of ANOVA
- Specify the factor or interaction to perform the comparison by selecting one of the radio buttons in the *Model Terms* frame

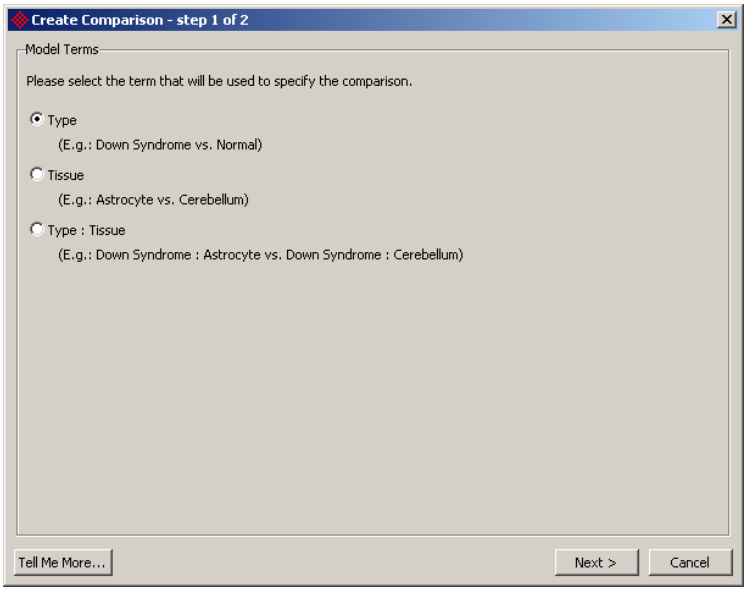

Create Comparison - step 1 of 2
Model Terms
Please select the term that will be used to specify the comparison.
Type
(E.g.: Down Syndrome vs. Normal)
Tissue
(E.g.: Astrocyte vs. Cerebellum)
Type : Tissue
(E.g.: Down Syndrome : Astrocyte vs. Down Syndrome : Cerebellum)
Tell Me More...
Next >
Cancel

Figure 9. 3: Create Comparison Dialog – Step 1

Configuring Comparison

The **Next >** button will lead you to configure the comparison (Figure 9. 4). You must specify two groups to compare. The left panel in the *Define Groups* lists all the levels (subgroups) of the selected factor or interaction.

Drag the levels in the left panel to *Group1* or *Group2*. Figure 9. 4 shows two brain tissues grouped together to compare to the heart tissue.

Using the Comparison Builder

You can drag and drop levels to assign levels to comparison groups. To do this, select one or more items from one group, hold the left mouse button down and drag them to the desired location, and then release the mouse. This will move the selected items from the original location to the new location.

When first brought up, all levels will be shown in the *Unassigned* group. To select multiple items, use <CTRL>, <Shift> and mouse click. Multiple selections can also be done by pressing down the left mouse button over a blank area (not over the text of an item) and dragging a bounding box (Figure 9. 5). All items that overlap with the bounding box will be selected. Clicking on the column header will select all items in the group. After the selection is made, start dragging.

Double clicking on an item in either *Group 1* or *Group 2* will move the item back to the *Unassigned* group. Double clicking on the column header of *Group 1* or *Group 2* will move all items currently in the group back to the *Unassigned* group.

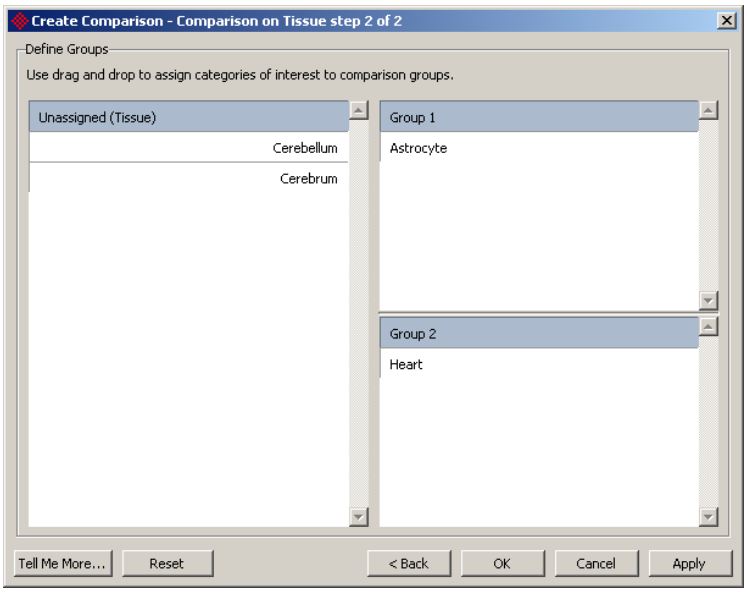


Figure 9. 4: Creating Comparison Dialog – Step 2 (Configuring comparison)

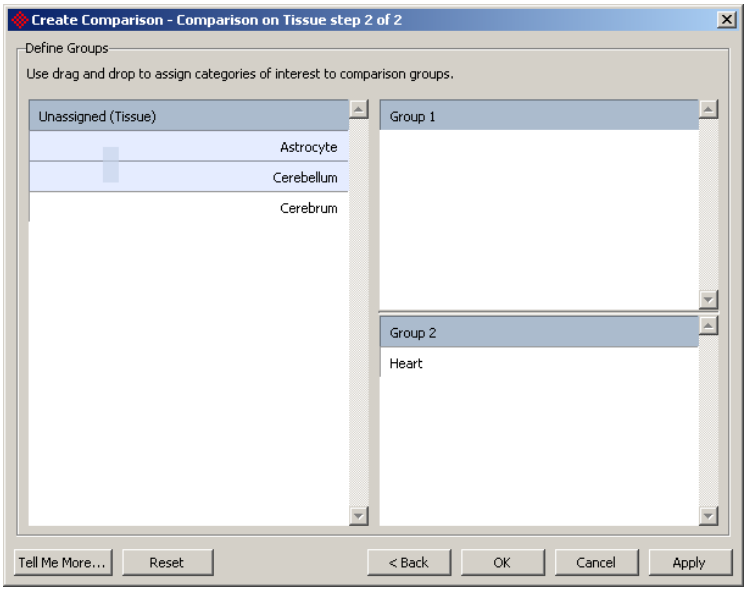


Figure 9. 5: Creating Comparison Dialog – Multiple Selection

Selecting **Apply** will add the specified contrast and stay on the same screen for more action.

Selecting **OK** will add the specified contrast and will send it to the *Comparison* list dialog (Figure 9. 6). *Non-Estimable* will be attached to the contrast name once the contrast is detected as non-estimable.

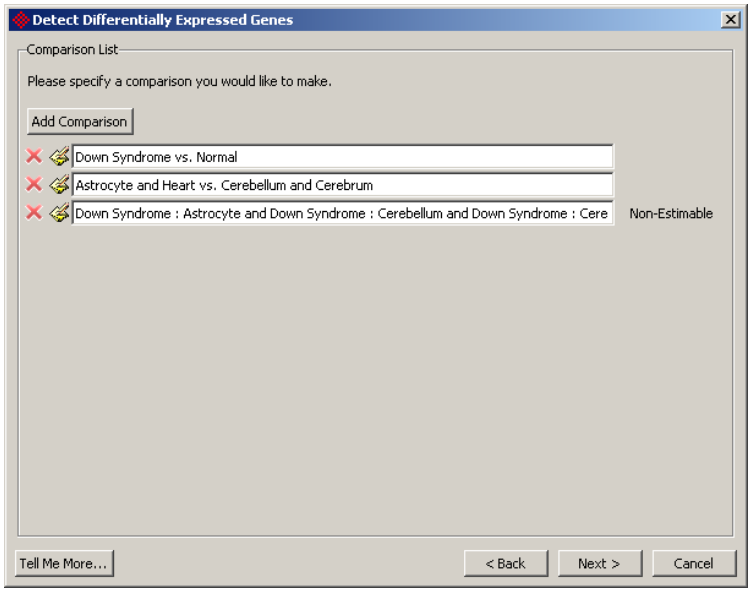

Detect Differentially Expressed Genes
Comparison List
Please specify a comparison you would like to make.
Add Comparison
Down Syndrome vs. Normal
Astrocyte and Heart vs. Cerebellum and Cerebrum
Down Syndrome : Astrocyte and Down Syndrome : Cerebellum and Down Syndrome : Cere
Non-Estimable
Tell Me More...
< Back
Next >
Cancel

Figure 9. 6: Viewing the Comparison List Dialog

Add Comparison

Selecting **Add Comparison** will add more comparisons.

Edit Comparison

will edit the corresponding comparison.

Remove Comparison

Selecting **Remove Comparison** will remove the corresponding comparison.

Rename Comparison

Editing the entry boxes will rename the corresponding comparison.

Running the Computations

Selecting **Next >** will perform the computation.

Selecting **Cancel** or the **<Esc>** key will close the dialog without doing any computation.

Selecting **< Back** will go back to the previous dialog to do more ANOVA configuration.

Running the computation includes two steps.

1. Assigning the batch effects
2. Running the ANOVA computation

Assigning the Batch Effects

The *Partek Express Estimate Gene Significance* will automatically assign batch effects for you based on the correlation and significance test.

The *Estimate Gene Significance* uses Cramer's V to do the correlation test between the batch and the categorical main factors and uses Pearson correlation coefficient to test the correlation between batch and numeric main factors. Experience shows that the batch probably needs to be excluded from the model if it has a quite strong correlation with the main effect or interaction since the strong correlation might indicate a confounding or nesting-nested relationship between the batch and the main factor or interaction. In this case, the effect of main factor will be stolen by the batch if the batch is included in the ANOVA model. When the computed value is larger than 0.8498, the batch will be excluded from the model.

The *Significance* test is used to test whether those batches that passed the correlation test are significant or not. *Estimate Gene Significance* uses model selection techniques to pick those batches that improve the model's adjusted R-Squares most.

Finally, Partek Express gives a summary page (Figure 9. 7) that shows what factors that have been selected as interest and what factors are recommended to be included as nuisance effects. Uncheck the batch effect if you do not want to include it in the model.

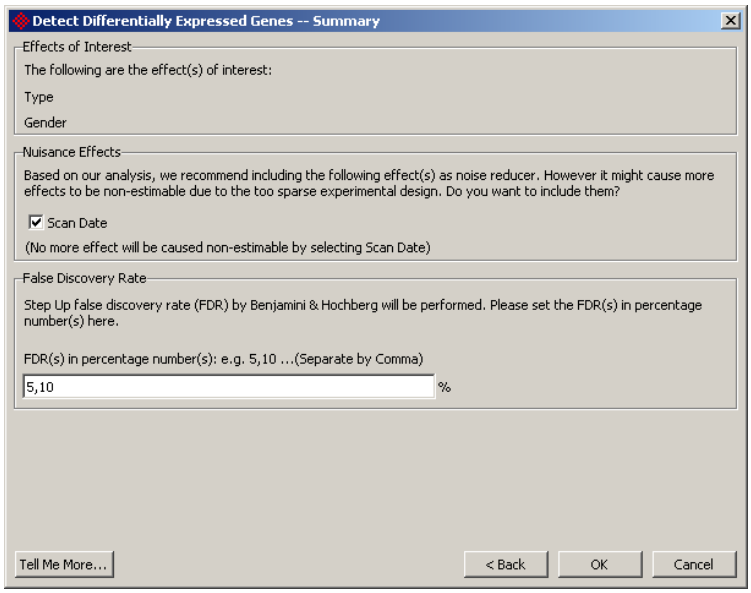

Detect Differentially Expressed Genes -- Summary
Effects of Interest
The following are the effect(s) of interest:
Type
Gender
Nuisance Effects
Based on our analysis, we recommend including the following effect(s) as noise reducer. However it might cause more effects to be non-estimable due to the too sparse experimental design. Do you want to include them?
Scan Date
(No more effect will be caused non-estimable by selecting Scan Date)
False Discovery Rate
Step Up false discovery rate (FDR) by Benjamini & Hochberg will be performed. Please set the FDR(s) in percentage number(s) here.
FDR(s) in percentage number(s): e.g. 5,10 ...(Separate by Comma)
5,10
%
Tell Me More...
< Back
OK
Cancel

Figure 9. 7: Estimate gene significance summary page

Setting the False Discovery Rate

A step-up multiple test correction will be automatically done for all the p-values produced by ANOVA. Setting the false discovery rate (Figure 9. 7) will produce a False Discovery Rate (FDR) report to the report tab when ANOVA is done. Please refer the *Multiple Test Correction for P-Values* section below to get more implementation details about step-up FDR.

Running the ANOVA computation

Selecting **OK** will perform the configured ANOVA computation.

Selecting **Cancel** or the **<Esc>** key will close the dialog without doing any computation.

Implementation Details

Sir Ronald Fisher first developed ANOVA* in 1925. Many intermediate statistical textbooks serve as an introduction to ANOVA (e.g. Steel and Torrie (1980) or Snedecor and Cochran (1980)). Scheffé (1959) is also a classic reference.

* ANOVA is a *parametric* test, it makes certain assumptions about the distribution of the response variable. The most important assumptions are that the data is normally distributed and that the variance is approximately equal between the groups (homogeneity of variance). Although ANOVA is most powerful when these assumptions are met, in many cases ANOVA is very robust to violations of these assumptions.

The Partek Express ANOVA can handle:

- a balanced and an unbalanced design
- random and fixed effects (mixed-model ANOVA), nested factors
- multi-number of categorical effects (multi-way ANOVA)
- numeric covariates (multi-way Analysis of Covariance, or ANCOVA)

Examples of each are provided below.

Example of a Balanced Experimental Design

A design is balanced when the number of samples is the same for each factor level. Consider the two factors, Treatment and Time. This is referred to as a 2X6 experiment design because Treatment has 2 levels and Time has 6 levels.

Factor	Levels
Treatment	Control, Treated
Time	T1, T2, T3, T4, T5, T6

Time vs. Treatment

Time\Treatment	Control	Treated	Total
T1	3	3	6
T2	3	3	6
T3	3	3	6
T4	3	3	6
T5	3	3	6
T6	3	3	6
Total	18	18	36

Figure 9. 8: An example of a balanced experimental design

Figure 9. 8 shows a balanced experimental design; in this case, a two-way crossed ANOVA. Every level of the factor Time occurs with every level of the factor Treatment. This is a balanced design because all of the levels of the two factors have the same number of samples (3).

An Example of Unbalanced Experimental Design

A design is unbalanced when the number of samples is not the same for each factor level. Below, in the Time and Treatment example, when a subject died at T6 (Time 6), the experiment became unbalanced (Figure 9. 9).

Time vs. Treatment

Time\Treatment	Control	Treated	Total
T1	3	3	6
T2	3	3	6
T3	3	3	6
T4	3	3	6
T5	3	3	6
T6	3	2	5
Total	18	17	35

Figure 9. 9: An example of an unbalanced experimental design

Missing Values & Missing Treatment Combinations

If the levels of all the factors are completely crossed, Type III sums of squares is used. However, if missing treatment combinations occur in any interaction, Type IV sums of squares is used.

A missing treatment combination occurs when one of the cells in the multi-way ANOVA table has no entries. If all three treated samples at time T6 are not available; then the Treated X T6 combination is missing (Figure 9. 10).

Time vs. Treatment

Time\Treatment	Control	Treated	Total
T1	3	3	6
T2	3	3	6
T3	3	3	6
T4	3	3	6
T5	3	3	6
T6	3	0	3
Total	18	15	33

Figure 9. 10: An example of a missing treatment combination

If an interaction corresponding to a treatment combination has no replication, *Partek Express* will automatically remove that interaction from the model. Therefore, when testing multiple response variables, the p-values of the removed interactions will be represented by question marks (“?”) to indicate that the value could not be computed.

Mixed Model ANOVA

To obtain estimates of variance components for mixed models, Partek Express uses the method of moments estimation (Eisenhart, 1947).

The method of moments is used to equate analysis of variance mean sum of squares to their expected values ($s=C\sigma^2$). S is a vector of the mean sum of squares, C is a matrix, and σ^2 is a vector of variance components. The estimates of σ^2 are $C^{-1}s$. However, the method of moments method can produce negative estimates.

Nested Factors

Most of the time, the grouping factor is nested with at least one of the main effects. In a two-way or a multi-way ANOVA, if each level of one factor occurs with each level of another factor, the factors are said to be "crossed". If, however, the levels of one factor only occur within a single level of another factor, then one factor is said to be "nested in" the other factor.

In the example below, there are two factors: Type and Subject ID. Type has 2 levels and Subject has 10 levels.

Factor	Levels
Type	Normal, TS21
Subject	1218, 1389, 1390, 1411, 1478, 1479, 1521, 1565, 748, 847

Notice in the *Crosstabulations* table (Figure 9. 11) that each level of Subject occurred within one level of Type. Therefore, Subject is nested within Type.

Subject vs. Type *

Subject\Type	Normal	TS21	Total
1218	0	3	3
1389	0	2	2
1390	4	0	4
1411	4	0	4
1478	0	4	4
1479	1	0	1
1521	4	0	4
1565	1	0	1
748	0	1	1
847	0	1	1
Total	14	11	25

* Subject is nested in Type.

Figure 9. 11: An example of a nested factor

Random vs. Fixed Effects

The grouping factor should be categorical (random). If it is not random, the *Estimate Gene Significance* procedure will automatically set it to random to do ANOVA and set it back after ANOVA is done.

Most factors in an ANOVA are fixed factors, i.e. the levels of that factor represent all the levels of interest. Examples of fixed factors include gender, race, strain, etc. However, in experiments that are more sophisticated, a factor can be a random effect, meaning the levels of the factor only represent a random sample of all of the levels of interest. Examples of random effects include subject and batch. Consider the example where one factor is *type* (with levels *normal, diseased*), and another

factor is *subject* (the subjects selected for the experiment). In this example, *type* is a fixed factor since the levels *normal* and *diseased* represent all conditions of interest. *Subject*, on the other hand, is a random effect since the subjects are only a random sample of all the levels of that factor.

Equations of the ANOVA Results

Contrast Equations

When contrasting the average of treatments A and B versus treatment C, the contrast equation is

$$\frac{1}{2}A+\frac{1}{2}B-1C=0$$

The ratio is calculated using the least square mean (LS Mean) of each term, thus the ratio for the contrast is given by:

$$\frac{\frac{1}{2}LSMean(A)+\frac{1}{2}LSMean(B)}{LSMean(C)}$$

A second example contrasts the average of treatments A, B, and C versus the average of treatments C and D. The contrast equation is

$$\frac{1}{3}A+\frac{1}{3}B+\frac{1}{3}C-\frac{1}{2}D-\frac{1}{2}E=0$$

The ratio for the contrast is given by:

$$\frac{\frac{1}{3}LSMean(A)+\frac{1}{3}LSMean(B)+\frac{1}{3}LSMean(C)}{\frac{1}{2}LSMean(D)+\frac{1}{2}LSMean(D)}$$

Fold changes are calculated in a similar fashion using LSMeans.

LS Mean and Geometric Mean

The LS Mean (Least Squares Mean) is calculated as the linear combination (sum) of the estimated means from a linear model (e.g. ANOVA, regression, etc). The LS mean is based on the factors specified in the model, thus, the LS mean is "model dependent" whereas arithmetic mean is "model independent". When the data results from a balanced experiment (same number of treatment combinations in each group), the arithmetic mean and LS mean are identical. In unbalanced data, the arithmetic mean and LS mean are different. In an unbalanced experiment, the LS means are preferred because they reflect the model being fit to the data.

Consider a simple unbalanced two-factor experiment containing a control group and a treated group, with unequal number of male and female animals in each group

(Figure 9. 12). The control group contains 4 females and 2 males, and the treated group contains 2 females and 5 males.

Treatment vs. Gender

Treatment\Gender	Female	Male	Total
Control	4	2	6
Treated	2	5	7
Total	6	7	13

Figure 9. 12: Unbalanced two-factor experiment crosstabulation

Using the arithmetic mean to estimate the means of the control and treated groups ignores the imbalance of male and female in the two groups and may be biased. For example, if you are estimating the effects of a gene's expression which lies on Y chromosome (females don't have a Y chromosome, thus they will have lower expression than males on this gene), the arithmetic mean would overestimate the mean in treated group since the treated group contains more males, and underestimate the control group since the control group contains more females (Figure 9. 13).

Arithmetic Mean

Treatment	N	Mean
Control	6	54.366540
Treated	7	255.856519

Least Squares Mean

Treatment	LSMean
Control	91.120693
Treated	208.601179

Figure 9. 13: Comparison of the arithmetic mean and LS mean in the control and treated group of gene's expression that lies on Y chromosome

The LS mean uses the estimates for both factors in the design, treatment, and gender, and adjusts the means for the treated and control groups to account for the imbalance in gender between the groups. The LS mean would produce a more accurate, unbiased estimate of the mean of the treated and control groups in this example (Figure 9. 13).

Data is often log transformed prior to doing statistical analysis in order to transform a multiplicative effect into an additive effect. However, scientists sometimes want to interpret effects as ratios, in which case log transformed data is inappropriate, since it has been converted from a multiplicative effect to an additive effect. Simply anti-logging the mean of logged data does not produce the mean of the un-logged data; however, it does produce the geometric mean of the un-logged data. Anti-logging a least squares mean produces a value that we call a "least squares geometric mean".

When a ratio is calculated based on LS means, the ratio of Group1 vs. Group2 is:

$$\frac{LSMean(Group1)}{LSMean(Group2)}$$

When a ratio is calculated based on least squares geometric means, the estimate of the LS means on logged data is first calculated for each group, and the difference of the LS means is then anti-logged using the same base:

$$a^{LSMean(Group1)-LSMean(Group2)}$$

"***a***" is the base the data is log transformed on. This is equivalent to calculating the ratio of the least squares geometric means for the two groups:

$$\frac{a^{LSMean(Group1)}}{a^{LSMean(Group2)}}$$

This ratio is more appropriate than the simple ratio of LS means in the case that analyses have been performed on logged data.

Multiple Test Correction for P-Values

A *p*-value is the probability that the observed values could have occurred by chance. It indicates the probability that one could obtain a test statistic that is as extreme as or more extreme than the observed one if the null hypothesis is true. *P*-values provide a sense of the strength of the evidence against the null hypothesis. The lower a *p*-value is, the stronger the evidence to reject the null hypothesis.

When multiple tests are performed, the probability of incorrectly rejecting a single null hypothesis ("false positive" or "Type I error") increases. There are several methods to correct Type I error for multiple tests. Partek Express only uses the most general one -- step up (Benjamini & Hochberg, 1995) method.

False Discovery Rate (FDR) – Step Up

False Discovery Rate is the proportion of false positives among all positives. In the step up method, there are n number of p-values; they are sorted by ascending order, and m represents the rank of a p-value. The calculation compares p-value*(n/m) with the specified significance level, and the cut-off p-value is the one that generates the last product that is less than the significance level.

Viewing the Effect Sizes Plots

The effect sizes tab presents either a bar chart or pie chart visualization to help display the importance and significance of experimental factors included in the analysis.

When using ANOVA to calculate p-values for effect sizes, an intermediate value, called the F-Ratio, is also calculated. F-ratio is a measure of the variance of the data explained by a factor relative to the unexplained variance or error.

To get an impression of the importance of each factor to the transcriptome overall, you can look at the mean of the factor's F-ratio across all genes on the bar chart or on the pie chart.

To switch between bar chart and pie chart views, select the appropriate sub tab within the effect sizes tab. Larger effect sizes indicate that the factor is more significant to the data.

In the bar chart visualization, each factor is represented by a vertical bar and is labeled with both the name of the factor and the height of the bar (Figure 9. 14). The vertical axis is the mean F-ratio available from the ANOVA table as described above. The error bar in the chart is always "1" so that the F-ratios are easier to interpret. Relative to this, taller bars represent more significant factors. Bars at, or near, error represent factors which are not significant to the transcriptome overall.

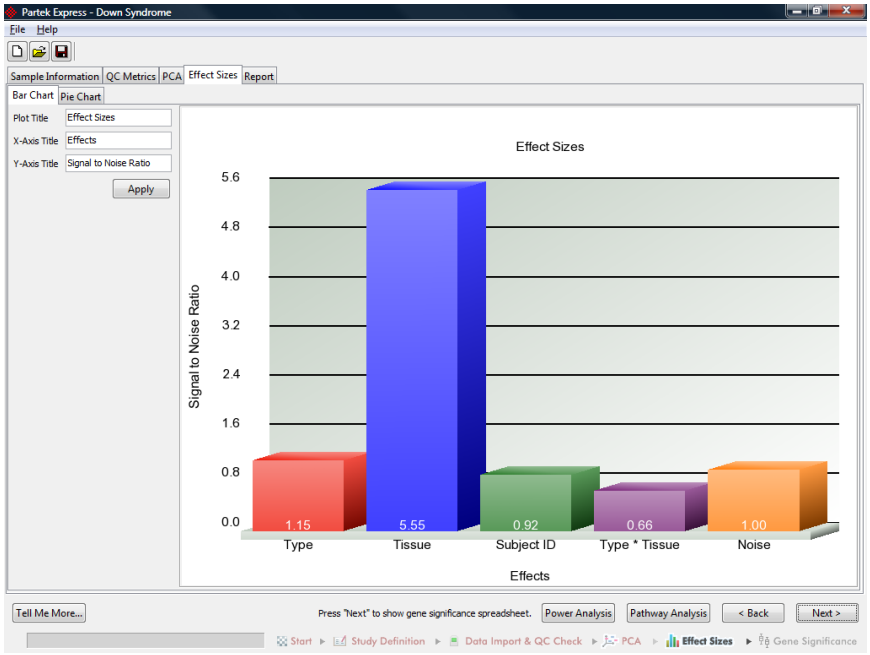


Figure 9. 14: Viewing the Effect Sizes Bar Chart

In the pie chart visualization, the magnitude of the F-ratios is used to generate a pie chart simplifying comparison of importance or effect between different factors (Figure 9. 15). Each section of the pie chart is labeled with the name of the factor and a percentage of the pie contained. Larger pieces of the pie are more significant, while factors at or near the size of the error slice represent factors which are not significant to the transcriptome overall.

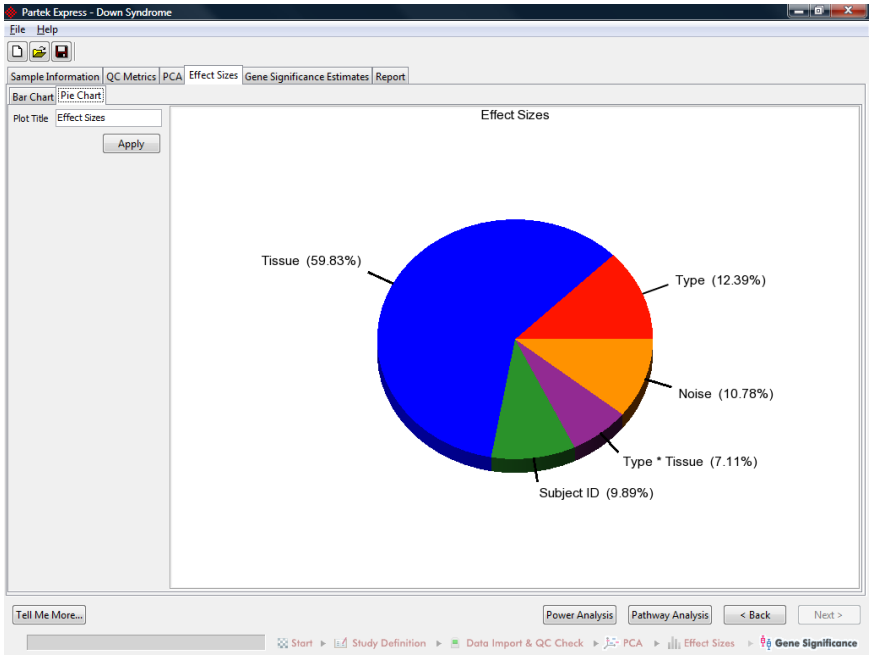


Figure 9. 15: Viewing the Effect Sizes Pie Chart

Configuring the Effect Sizes Plots

The title or axis labels of the effect size visualizations can be set on the left of the visualization. To make changes, type in the desired title and axis labels and select **Apply**.

References

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics, 3: 1-21.

Tamhane, Ajit C., & Dunlop, Dorothy D. (2000). Statistics and Data Analysis from Elementary to Intermediate. Prentice Hall. Pages 473-474.

Thompson, W.A., Jr (1962). The Problem of Negative Estimates of Variance Components. Ann. Math. Stat. 33: 273-289.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, JRSS, B, 57, 289-300.

Partek® Express™: Viewing Gene Significance

Introduction

Partek® Express™ estimates Gene Significance using Analysis of Variance (ANOVA). For more information about ANOVA, please refer to the *Effect Sizes* chapter.

Viewing Gene Significance Results

The ANOVA results will be summarized in a spreadsheet and a dot plot (Figure 10. 1) which are located in the *Gene Significance Estimates* tab.

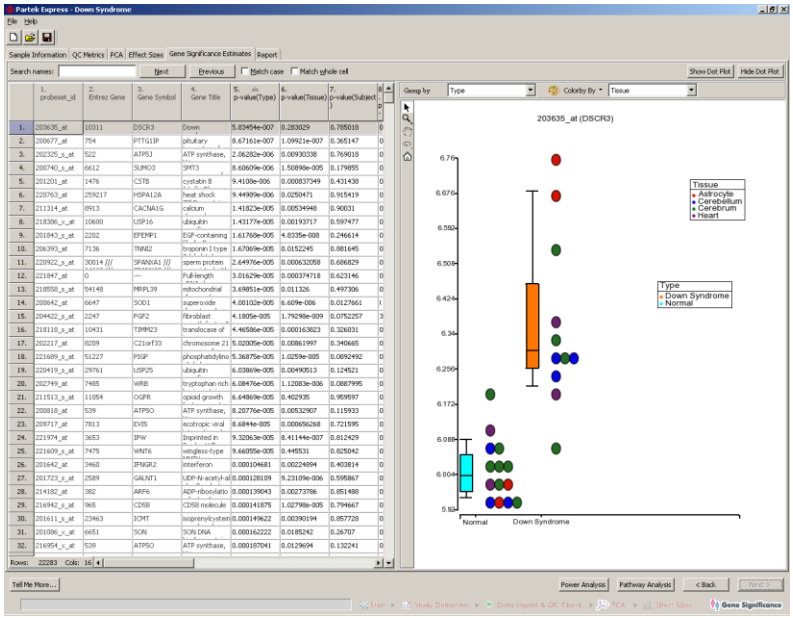


Figure 10. 1: Estimate Significance Gene Results

Searching for Genes

On top of the *ANOVA Result Spreadsheet*, a gene name search panel is provided. To search for a gene name, type it into the entry box and choose **Next** or **Previous**. If **Next** is selected, the search will begin at the currently active cell and go downwards. If **Previous** is selected, the search will begin at the currently active cell and go upwards. Two options, **Match case** and **Match whole cell** are available for search. To enable an option, click the checkbox next to it.

If a match is found, the matched cell will be activated and the matching row will be highlighted. You can continue searching down or up by clicking **Next** or **Previous**.

Result Spreadsheet

The analysis is performed on imported intensities for each gene. Each row of the *Results Spreadsheet* corresponds to one gene.

One of the most critical pieces of information contained in the *Gene Significance* table is the p-value per gene per categorical variable. A p-value is a test statistic (between zero and one) used to rank significance of results of starting with the null hypothesis that a gene is similarly expressed across conditions. Stated slightly differently and somewhat simplistically, the smaller the p-value for a given gene, the more likely that the gene shows differential express across the given categorical variables.

Each biological factor included in the ANOVA model will produce one additional column in the *Gene Significance* table. Each pair-wise comparison included in the ANOVA will add three additional columns into the table.

By default, the genes are sorted by the p-values of the first factor of interest. The most significant gene is on the first row. To sort by a different column, simply right click on the column heading and select **Sort Ascending** or **Sort Descending** in the pop-up menu.

Dot Plot

Dot plot can be viewed gene by gene just by clicking the row label corresponding to the response variable. It shows original intensities grouped by the factor in the **Group by** dropdown list and colored by the factor in the **Color by** dropdown list.

In the dot plot, each dot is an individual sample data point. The X-Axis represents the different types and the Y-Axis displays the log2 expression level of the gene. When data is in log2 space, it is important to remember that the scale is typically between zero and 16, and that any increment change of one represents a twofold change in abundance. So if a gene changes from 6 in one condition to 8 in another condition, that represents a fourfold change between the two conditions.

References

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics, 3: 1-21.

Tamhane, Ajit C., & Dunlop, Dorothy D. (2000). Statistics and Data Analysis from Elementary to Intermediate. Prentice Hall. Pages 473-474.

Thompson, W.A., Jr (1962). The Problem of Negative Estimates of Variance Components. Ann. Math. Stat. 33: 273-289.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, JRSS, B, 57, 289-300.

Partek® Express™: Power Analysis

Introduction

The Partek® Express™ Power Analysis procedure conducts prospective analysis, which is used to:

- Determine the minimum sample size to achieve adequate power on a given fold change
- Determine what fold change could be acquired on the given sample size to achieve the specified power

Implementation Details

Input for *Power Analysis* includes:

- Experimental design
- Statistical model (ANOVA)
- Comparison (contrast) on which to do power analysis
- Effect size (fold change)
- Sample size
- Significance level (alpha)
- Power (1-beta)

Partek Express Power Analysis obtains the experimental design, statistical model (ANOVA) and comparison (contrast) from the current study. These three parameters are already decided by previous steps in *Partek Express* before doing power analysis. Please refer the *Estimate Gene Significance* chapter to get more information about how to configure an ANOVA model, how to set-up comparisons, as well as implementation details.

Let Y be the response vector, X be the design matrix, β be the model parameter vector, so the underlying function for the ANOVA model can be written in the form of $Y = \beta X + \varepsilon$ where ε is the error term which is normally and independently distributed with mean 0 and standard deviation σ. *Comparison (contrast)* was set in the *Estimate Gene Significance (ANOVA)* step to test the null hypothesis $H_0 : \mathrm{L}\beta = 0$ where L is the contrast matrix.

For the four parameters like effect size, sample size, significance level and power, each can be obtained by solving the following power analysis formulas when fixing the other three.

Power Analysis Formulas

$power = P(F(r_L, N - r_x, \lambda) \geq F_{1-\alpha}(r_L, N - r_x))$ (Muller and Peterson 1984)

Where r_L is the rank of contrast L, r_x is the rank of design matrix *X*, *N* is the total sample size, α is the significance level and λ is the non-central parameter of F statistic under alternative hypothesis $H_A : L\beta \neq 0$.

$\lambda = N(L\beta)'(L(\ddot{X}'diag(w)\ddot{X})^{-1}L')^{-1}(L\beta)\sigma^{-2}$

Where $\ddot{X}$ is composed of the unique rows of design matrix *X*, *w* is a vector of weights which reflect the proportion of each unique row in the whole design matrix *X*. σ is the ANOVA model standard deviation.

Configuring the Power Analysis Dialog

The *Partek Express* embedded workflow will introduce you to the *Power Analysis* step automatically by consistently selecting the **Next** button on the main window. The *Power Analysis* step is after the *View Effect Sizes* step: **Start > Study Definition > Data Import & QC Check > PCA > Effect Sizes > Power Analysis.**

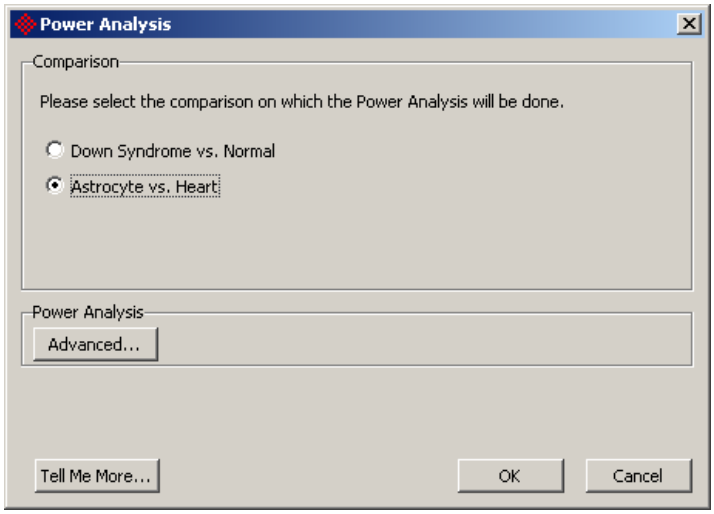


Figure 11. 1: Configuring Power Analysis

Selecting Comparison

All the comparisons that were set in the *Estimate Gene Significance* step will be shown in the *Power Analysis* dialog (Figure 11. 1); however, only one comparison can be selected to do power analysis at a time. To specify one comparison, just click one of the radio buttons in the *Comparison* frame. If no comparison was set in *Estimate Gene Significance* step, the *Power Analysis* dialog will not be invoked.

Configuring the Effect Size

Selecting the **Advanced...** button in the *Power Analysis* frame will open the *Power Analysis Configuration* dialog (Figure 11. 2) to configure the parameters of effect size, sample size, significance, and power.

Specify the range and step size for effect size in this dialog so that the *Power Analysis* will produce the minimum sample sizes (the newly produced sample size is supposed to be assigned to each comparison group with the same proportion as the original dataset) required to achieve each of the specified effect sizes, respectively. Effect size (fold change here) must be greater than or equal to one. Decreasing the effect size will probably require more samples. For better viewing, 10 points of effect size can be accommodated in the specified range by the specified step size.

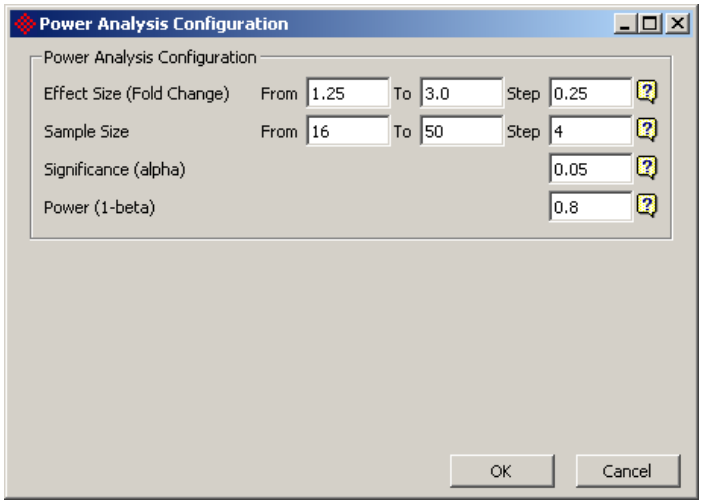


Figure 11. 2: Configuring Power Analysis

Configuring the Sample Size

Specify the range and step size for the sample size so that the *Power Analysis* will produce a fold change that is set by the given sample sizes. Sample size should be larger than model's degree of freedom. For better viewing, 10 points of sample size can be accommodated in the specified range by the specified step size.

Configuring the Significance

The significance level is the probability to reject the null hypothesis when the null hypothesis is actually true. A commonly used significance level of 0.1 is set as the default. The range for significance level is between 0 and 1. Decreasing the significance level will probably require more samples to achieve the same fold change.

Configuring the Power

The power level is the probability to reject the null hypothesis when the null hypothesis is actually false. A commonly used power of 0.8 is set as the default. The range for power is between 0 and 1. Increasing the power will probably require more samples to achieve the same fold change.

Saving the Power Analysis Configuration

Selecting **OK** in the *Power Analysis Configuration* dialog (Figure 11. 2) will save all the parameters configured and dismiss the dialog; selecting **Cancel** will close the dialog without saving.

Running the Power Analysis

Selecting **OK** in the *Power Analysis* dialog (Figure 11. 1) will perform the configured power analysis and dismiss the dialog; selecting **Cancel** will close the dialog without doing any computation.

Visualizing the Data: Box plot

The box plot provides a way to graphically view the numeric data through five numbers in summary. The five numbers, 10^{th} percentile, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile and 90^{th} percentile of the power analysis, result in the gene level. *Partek Express* Power Analysis will generate two box plots, *Fold Change to Sample Size* and *Sample Size to Fold Change*. These two box plots can be invoked by selecting the radio button on the *Power Analysis* tab in the *Partek Express* main window.

Box Plot: Fold Change to Sample Size

The *Fold Change to Sample Size* box plot indicates the sample size (in Y axis) to achieve the adequate power of the given fold change (in X axis).

Figure 11. 3: Box plot of Fold Change to Sample Size

Note: Y axis tick marks are in log (base 2) scale. The current study sample size is marked with a blue reference line. Moving the mouse over a box-whisker will show

a more detailed sample size report. In the examples shown in figure 3, to detect 50% of genes with a fold change of at least 1.75 would require 20.85 (round up to 21) samples.

Note: Power Analysis for a specific fold change assumes the proportion of samples in each category is similar to that of the existing samples. Table 11. 1shows the number of samples needed for a fold change of at least 1.75.

# of Samples	Percent of Genes	Fold Change
18.70	10%	1.75
19.44	25%	1.75
20.85	50%	1.75
24.17	75%	1.75
32.32	90%	1.75

Table 11. 1: Viewing the number of samples needed to achieve a fold change of at least 1.75

Box Plot: Sample Size to Fold Change

The *Sample Size to Fold Change* box plot shows what fold change (in X axis) could be acquired on the given sample size (in Y axis).

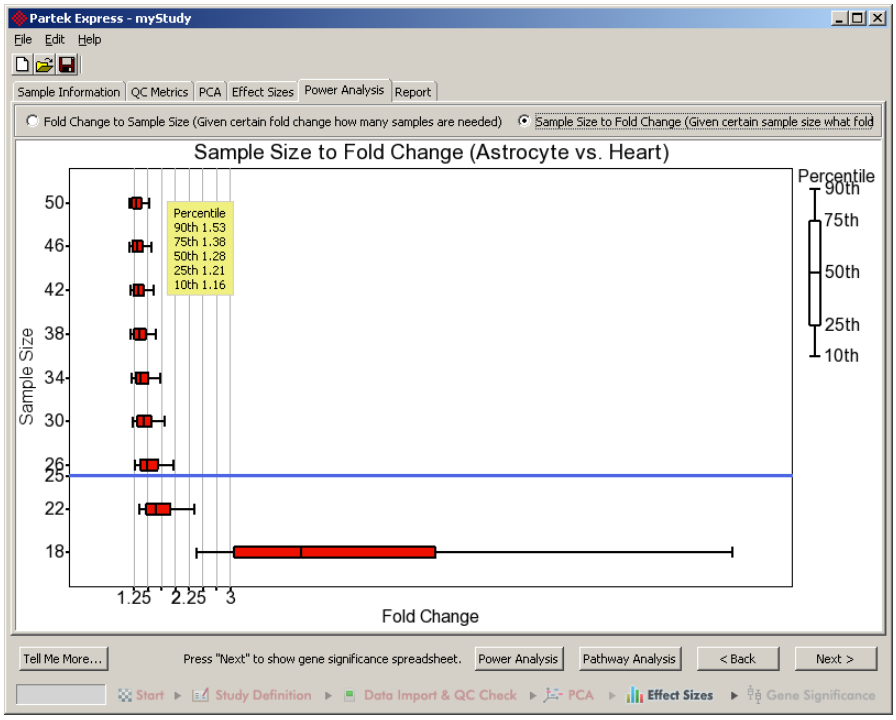


Figure 11. 4: Box plot of Sample Size to Fold Change

The blue line marks the number of samples used in the analysis of the study. Moving the mouse over a box-whisker will bring up the detailed fold change report for the respective samples size. In the example shown in Figure 11. 4, using 50 samples in the study would detect 50% of genes with a fold change of at least 1.28.

Note: Power Analysis for a specific sample size assumes the proportion of samples in each category is similar to that of the existing samples. Table 11. 2 shows the fold change of 50 samples at varying percentages.

# of Samples	Percent of Genes	Fold Change
50	10%	1.16
50	25%	1.21
50	50%	1.28
50	75%	1.38
50	90%	1.53

Table 11. 2: Viewing the fold change of 50 samples at differing percentages

References

Muller, K.E. and Peterson, B. L. (1984), Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics and Data Analysis*, 2: 143-158.

Muller, K.E. and Benignus, V.A. (1992), Increasing scientific power with statistical power. *Neurotoxicology and Teratology*, 14: 211-219.

Muller, K.E., LaVange, L.M., Ramey, S.L. and Ramey, C.T. (1992), Power calculations for general multivariate models including repeated measures applications. *Journal of the American Statistical Association*, 87: 1209-1226.

Partek® Express™: Report

Introduction

The Partek® Express™ Report tab records every step that has been done in the current study. It can be used to:

- Determine the data set used in the study
- Determine what analysis was done and how it was done
- Find the reference paper for each analysis performed

Description of Report

The report provides step by step information about the actions taken in the Partek Express workflow. The main steps recorded are: *Create Study, Import, PCA, ANOVA* and *Power Analysis*. The reference papers that are related to each of the above steps are all recorded in the *References* section.

Create Study

From the *Report* tab, you can find the name of the study, the time on which the study was created, and the user who created the study.

Import

The report of the *Import* step will tell you which files were imported, where to find the library files, and what algorithm was used to do the importing. The algorithm Partek Express uses to import CEL or CHP files is RMA.

RMA

The Partek Express implementation of RMA is tuned for speed and decreased memory usage. There are four steps involved in the RMA importing method; only perfect match (PM) probe values are used in this method:

- Background correction is used on the PM values
- Quantile normalization is used across all the chips in the experiment
- Log (base 2) transforms the data, and if the data values are <= 0, then they will be marked as *missing*
- Median polish summarization is used to give a robust signal for each gene. Note: Median polish might give the same summarized values for all/most samples if your sample size is very small. When the sample size is <=4, mean summarization is used

PCA

The correlation method used to draw a principal components analysis (PCA) plot is recorded in the *PCA* section.

ANOVA

The report for the *ANOVA* step will help you find the file on which ANOVA was performed. It will tell you what method was used as well as what model was created to do the ANOVA. A Step Up false discovery rate (FDR) report will be produced after the ANOVA is done.

Power Analysis

The report for the *Power Analysis* step remembers all the parameters configured in the Power Analysis dialog. The parameters include the comparison on which the Power Analysis was performed, effect size, sample size, significance level and the power specified to do power analysis.

References

All reference papers related to the performed steps, above, will be shown in the *References* section.

Partek® Express™: Pathway Analysis

Introduction

The Partek Express Pathway Analysis is used to:

- Dump useful information from the ANOVA results
- Provide an interface to launch Ariadne Pathway Studio® Explore or Ingenuity® IPA® to do the Pathway Analysis

Please note: you need to have Ariadne Pathway Studio® Explore or Ingenuity® IPA® installed on the same computer as Partek® Express™ is to do the Pathway Analysis step.

Configuring the Pathway Analysis Dialog for Ariadne Pathway Studio Explore

The Partek Express embedded workflow will introduce you to the *Pathway Analysis* step automatically by selecting the **Next** button on the main window. The *Pathway Analysis* step is after the *View Effect Sizes* step: **Start > Study Definition > Data Import & QC Check > PCA > Effect Sizes > Pathway Analysis.**

The *Ariadne* tab in the *Pathway Analysis* dialog will configure and launch Ariadne Pathway Studio Explore (Figure 13. 1).

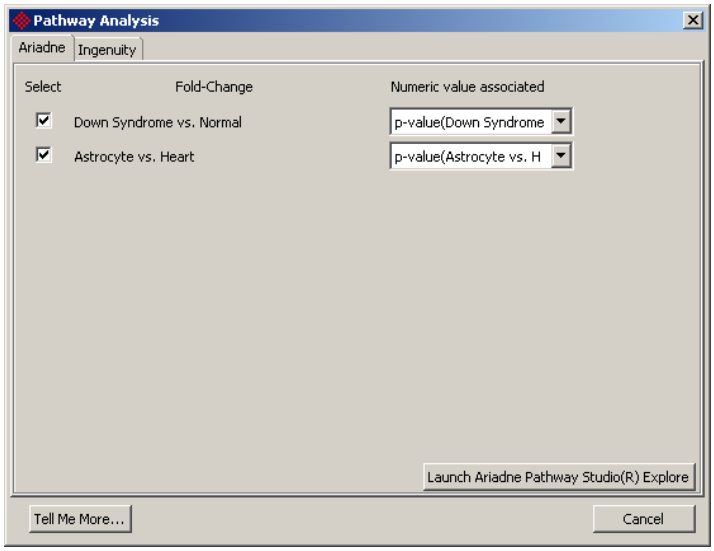


Figure 13. 1: Configuring Ariadne Pathway Analysis

Selecting Ratio

Estimate Gene Significance step generated a ratio column for each comparison in the result spreadsheet. All those comparisons' name will be shown in the *Pathway*

Analysis dialog (Figure 13. 1) except that whose ratio column is all composed of question marks (non estimable values). To select one ratio, just check the check button in front of the desired comparison name.

Selecting Numeric Value Associated

The p-value column with the same comparison name as ratio is set as the default *Numeric value associated*. If this column is all composed of question marks, any numeric column with the same comparison name as ratio will be set as default. To change the *Numeric value associated,* click the dropdown list box and select one from the dropdown list.

Running the Pathway Analysis with Ariadne Pathway Studio Explore

- Selecting **Launch Ariadne Pathway Studio® Explore** will output the *ratio* and *Numeric value associated* for each comparison selected in the Pathway Analysis dialog (Figure 13. 1) and launch Explore to do pathway analysis
- Selecting **Cancel** will close the dialog without doing pathway analysis
- Selecting **Tell Me More…** will lead you to this Pathway Analysis manual

Configuring the Pathway Analysis Dialog for Ingenuity IPA

The *Ingenuity* tab in the *Pathway Analysis* dialog will launch Ingenuity Pathway Analysis (Figure 13. 2).

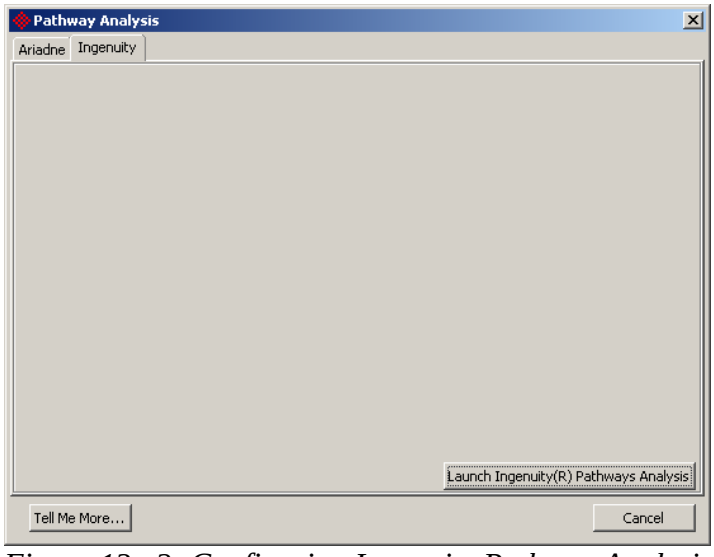


Figure 13. 2: Configuring Ingenuity Pathway Analysis

- Selecting the **Launch Ingenuity® Pathways Analysis** will invoke the default browser. Selecting **Send to IPA** will launch Ingenuity IPA for the pathway analysis (Figure 13. 3)
- Selecting **Cancel** will close the dialog without doing pathway analysis
- Selecting **Tell Me More…** will lead you to this Pathway Analysis manual

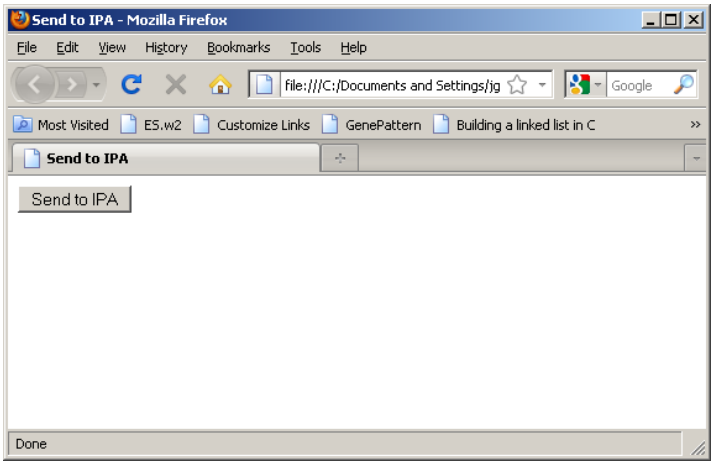

Send to IPA - Mozilla Firefox
File Edit View History Bookmarks Tools Help
file:///C:/Documents and Settings/jg
Google
Most Visited ES.w2 Customize Links GenePattern Building a linked list in C
Send to IPA
Send to IPA
Done

Figure 13. 3: Launching Ingenuity Pathway Analysis

Partek® Express™: The Main Menu

In this chapter, you will find information about the Partek® Express™ menus.

The File Menu

New Study

Creates a new study. If another study is currently open in Partek Express, it will close the current study first. You will be prompted to save the current study if there are pending unsaved changes.

Load Study

Allows you to choose a study file to open. If another study is currently open in Partek Express, it will close the current study first. You will be prompted to save the current study if there are pending unsaved changes.

Save Study

Saves the current study.

Save Study As…

Saves the current study with a different name.

Zip Study

Allows you to specify a .zip file to package the current study and associated data files (.CEL/.CHP and .ARR files) into (Figure 14. 1). Check the appropriate checkboxes to include the data files

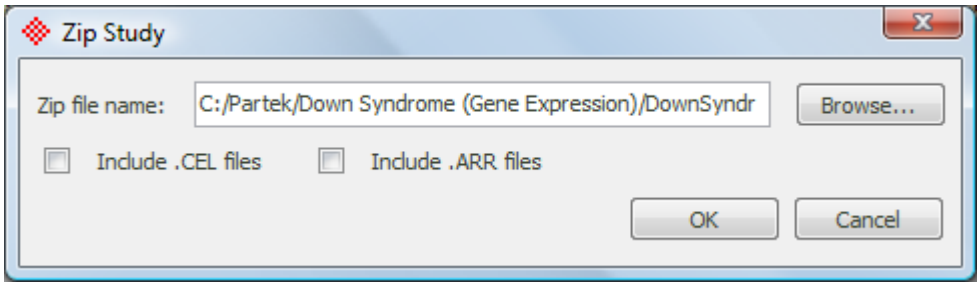


Figure 14. 1: Viewing the Zip Study Dialog

If there are pending unsaved changes, you will be prompted to save the study before continuing.

Close Study

Closes the current study.

Recent Studies

Opens a recent study.

Save Image As…

Saves the current image in the viewer. This menu is only available for some plots/charts, such as the QC Metrics Plot, PCA Plot, Effect Sizes Bar Chart, Effect Sizes Pie Chart, Gene Significance Estimates Dot Plot, and Power Analysis Box Whisker Plot.

Export

Under the **Export** cascade menu there are the following items.

Sample Information (Tab Delimited)

Saves the sample information as a tab delimited text file.

Sample Information (.ARR)

Exports the sample information into Affymetrix .ARR file(s). This operation requires that the .CEL or .CHP files specified in the sample information exist. The .ARR files will be exported to the same folder where the data files are located. If .ARR files already exist, you will be prompted to overwrite.

Intensities (Genes on Rows, Excel compatible)

Exports sample intensity values as a tab delimited file. Each column of the file will contain one sample. The file format is compatible with Microsoft Excel. This operation is only available after *Data Import & QC Check*.

Intensities (Samples on Rows, Partek GS compatible)

Saves the sample intensities to two files (e.g. xyz and xyz.fmt). xyz will have the intensities, and xyz.fmt will have the format information. You can use Partek® Genomics Suite™ (Partek GS), to open xyz.fmt. Note: you will need both files (xyz and xyz.fmt) in order for Partek GS to open.

Gene significance Estimates (Tab Delimited, Genes on Rows)

Exports the gene significance estimates spreadsheet as tab delimited text files with one row representing one gene. This file is Excel compatible.

Report

Exports the study information under the *Report* tab to a text file.

Manage Library Path

Brings up a folder browser (Figure 14. 2). Here the user specified folder will be used to search for library files when importing data; this will also be the directory Partek Express automatically downloads library files to.

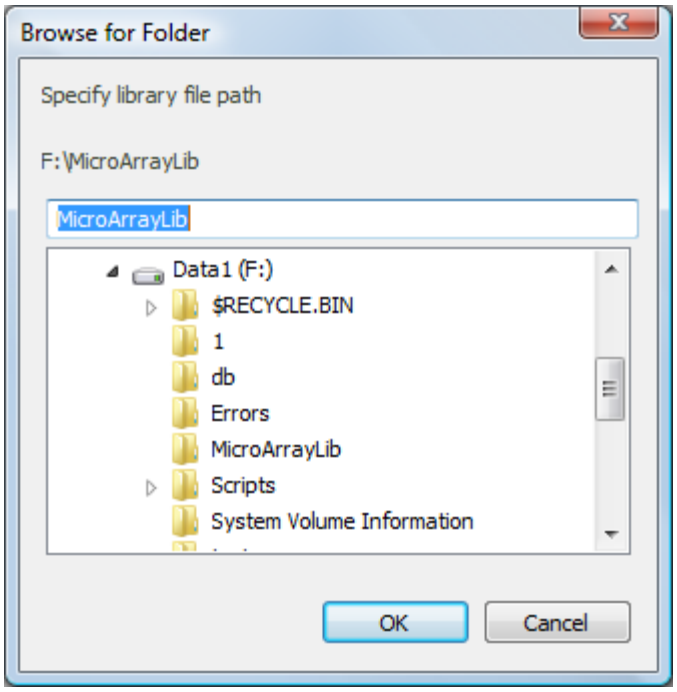


Figure 14. 2: Viewing the Browse for Folder Dialog

Manage Library Files

Invokes the file manager (Figure 14. 3), which allows for the update of annotation and library files. For example, in Figure 14. 3, checking the *Update Available* button, and then selecting the *Download* button will get the latest *HG-U133A Probeset Annotation* file.

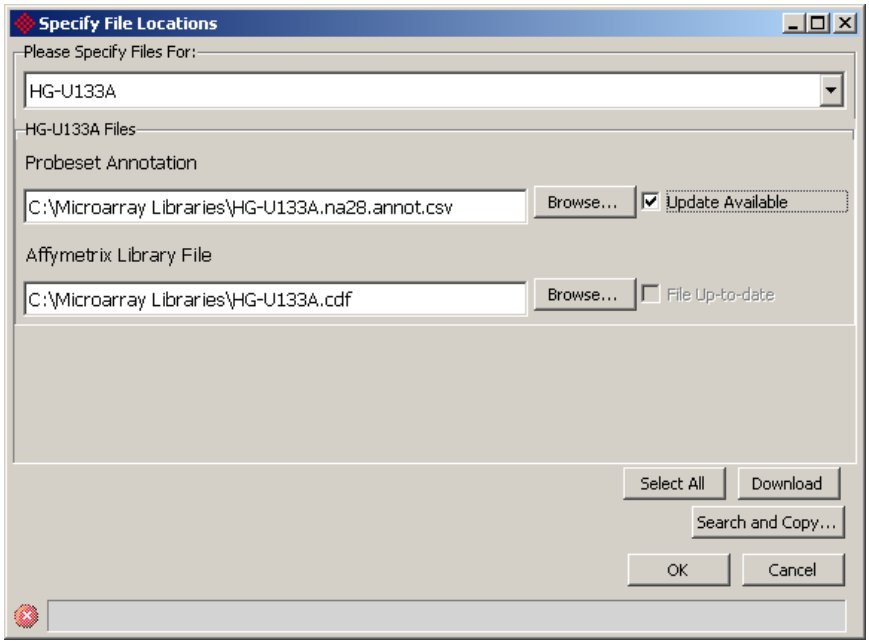


Figure 14. 3: Viewing the File Manager Dialog

Exit

Closes the application.

The Edit Menu

Plot Fonts Configuration

Invokes the *Plot Fonts Configuration* dialog (Figure 14. 4). Here you can specify the font size for the *Title, Axis, Axis Title, Legend,* and *Label,* and specify the *Plot Font.*

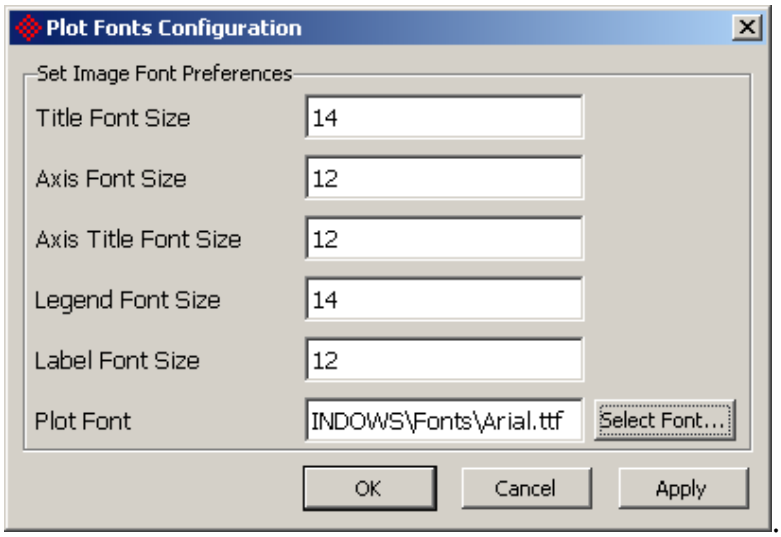


Figure 14. 4: Plot Fonts Configuration Dialog

The Help Menu

On-line Tutorial

Launches a web browser, which will show the tutorial at: http://www.partek.com/~devel/PartekExpressDownSyndrome_tutorial.pdf

User's Manual

Invokes the Partek Express User's Manual.

License Information…

Displays license information (Figure 14. 5).

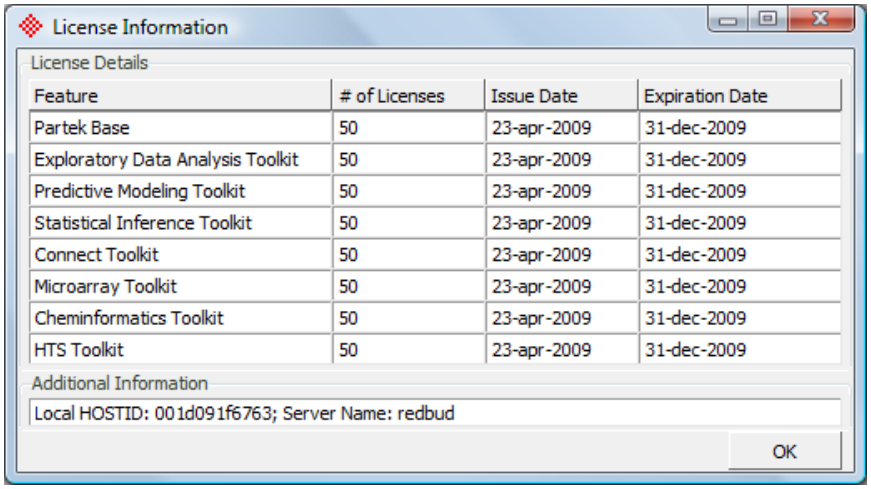


Figure 14. 5: Viewing License Information

Graphics Information…

Displays graphics information.

Check for Updates

Displays the Partek Express update webpage in a web browser.

About Partek Express

Displays information about the software.

Copyright © 2010 by Partek Incorporated. All Rights Reserved. Reproduction of this material without express written consent from Partek Incorporated is strictly prohibited.